ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-798

A Global Regression Discontinuity Design: Theory and Application to Grade Retention Policies

Isaac M. Opper
RAND Corporation

Umut Özek
RAND Corporation

We propose a novel estimator for use in a fuzzy regression discontinuity setting. The estimator can be thought of as extrapolating the traditional fuzzy regression discontinuity estimate or as an observational study that adjusts for endogenous selection into treatment using information at the discontinuity. We show that it can be motivated as being the least complex model consistent with the data or as an estimator that is preferable to both a traditional regression discontinuity design and an observational study. We further show theoretically that no other estimators consistently generate better estimates than our proposed estimator. We then use this approach to examine the effects of early grade retention beyond the compliers around the retention cutoff. We show that the benefits of early grade retention policies are larger for students with lower baseline achievement and smaller for low-performing students who are exempt from retention. These findings imply that (1) the benefits of early grade retention policies are larger than have been estimated using traditional fuzzy regression discontinuity designs and (2) retaining additional students would have a limited effect on student outcomes.

VERSION: June 2023

Suggested citation: Opper, Isaac M., and Umut Özek. (2023). A Global Regression Discontinuity Design: Theory and Application to Grade Retention Policies. (EdWorkingPaper: 23-798). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/hq2t-7x64

A Global Regression Discontinuity Design: Theory and Application to Grade Retention Policies*

Isaac M. Opper† Umut Özek‡

June 26, 2023

Abstract

We propose a novel estimator for use in a fuzzy regression discontinuity setting. The estimator can be thought of as extrapolating the traditional fuzzy regression discontinuity estimate or as an observational study that adjusts for endogenous selection into treatment using information at the discontinuity. We show that it can be motivated as being the least complex model consistent with the data or as an estimator that is preferable to both a traditional regression discontinuity design and an observational study. We further show theoretically that no other estimators consistently generate better estimates than our proposed estimator. We then use this approach to examine the effects of early grade retention beyond the compliers around the retention cutoff. We show that the benefits of early grade retention policies are larger for students with lower baseline achievement and smaller for low-performing students who are exempt from retention. These findings imply that (1) the benefits of early grade retention policies are larger than have been estimated using traditional fuzzy regression discontinuity designs and (2) retaining additional students would have a limited effect on student outcomes.

*We thank Christine Mulhern for the incredibly helpful conversations about the method. We also thank conference participants at 2023 AEFP for their thoughtful comments and the anonymous school district for providing the data used in the analysis. Code to implement the approach is currently undergoing RAND's Quality Assurance and will be made publicly available when that process is complete; if one wants to view the code before it is made publicly available, feel free to email iopper@rand.org. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D00008. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

† RAND Corporation. Email: iopper@rand.org.

‡ RAND Corporation. Email: uozek@rand.org.

I Introduction

Regression discontinuity (RD) designs have become increasingly popular in empirical research over the past three decades (Cook (2008), Abadie and Cattaneo (2018), Brodeur et al. (2020), Boon et al. (2021), and Wuepper and Finger (2023)). This framework leverages plausibly exogenous discontinuities in treatment likelihood at predetermined cutoffs to identify the causal effect of the treatment (Imbens and Lemieux (2008), Abadie and Cattaneo (2018)). When the discontinuity in treatment likelihood is fuzzy – i.e., some individuals on the treatment side of the cutoff do not receive treatment while some individuals on the other side receive treatment – a common approach is to use an instrumental variables (IV) design where being on the treatment side of the cutoff is used as an instrument for receiving treatment (Imbens and Lemieux (2008), Abadie and Cattaneo (2018)). While such fuzzy designs generally provide compelling evidence of the treatment effect, the IV estimator yields a local average treatment effect (LATE) in two ways.

First, the estimated effects only apply to individuals for whom being on the treatment side of the cutoff determines treatment status (Bertanha and Imbens (2020)). Second, similar to sharp RD designs with perfect compliance, it is hard to generalize these estimates to individuals identified for treatment who are away from the cutoff (Angrist and Rokkanen (2015), Cattaneo et al. (2021), Dong and Lewbel (2015)). Yet, understanding treatment effects beyond compliers at the cutoff is important from a public policy perspective. For example, moving beyond LATE is necessary if one wants to (1) assess how increasing compliance among those identified for treatment may influence the effectiveness of the policy; (2) understand whether exemptions often incorporated into public policies indeed identify individuals less likely to benefit from treatment; or (3) examine the effect of the treatment on individuals away from the cutoff who typically have higher needs (e.g., educational), which is essential to assess the overall benefits of the policy.

This paper addresses the locality issue by introducing a new estimator for use in a fuzzy regression discontinuity setting that generates global treatment effect estimates. The estimator jointly models the two potential outcomes and selection into treatment. We then greatly restrict the set of potential control functions to ensure that our estimator converges to a unique marginal treatment effect (MTE) function.

Although identification is obtained by restricting the set of potential control func-

tions, we show that the estimator can be easily interpreted and motivated even in cases where the true MTE function does not lie in this restricted set. For example, we show that the estimator can be thought of as an extrapolation of the traditional fuzzy regression discontinuity estimate, where the initial extrapolation to the non-compliers at the threshold is done using existing approaches usually employed in the RCT setting (e.g., Brinch et al. (2017); Kowalski (forthcoming)) and extrapolation away from the threshold is done using the assumption that the amount of endogenous selection stays constant. Similarly the estimator can also be interpreted as starting with an observational study and then adjusting for bias using information at the discontinuity in a similar fashion as Bertanha and Imbens (2020).

We also highlight two additional ways the approach can be motivated. First, we show that the resulting estimate is the least complex model that is consistent with the data observed. We then develop a Bayesian hierarchical model to contrast our estimator with two alternative approaches: estimating the ATE assuming no endogenous selection into treatment (e.g., ignoring the discontinuity) and estimating the average treatment effect (ATE) using the local estimate at the discontinuity. We then show that our estimator dominates these two alternatives, in that it has a lower mean-squared error regardless of what hyperparameters are employed in the Bayesian hierarchical model. We conclude our theoretical analysis by showing that no other estimator dominates our proposed estimator, i.e., no alternative estimator has a lower mean-squared error for all plausible hyperparameters.

We then use this estimator to examine the broader effects of early grade retention policies in the United States. This application is important for two reasons. First, this exercise has important implications for education policy in the United States: as of 2020, about half of all states and the District of Columbia require or encourage school districts to retain third-grade students who lag behind based on their third-grade reading scores. There is growing literature examining the effects of these policies using RD designs[1], yet we know very little about their effects on students away from the retention cutoff who have lower initial achievement.

Second, several early grade retention policies include "exemptions" to test score thresholds, such as for students who have disabilities, who are recent English learners,

[1]For example, see Greene and Winters (2007), Winters and Greene (2012), Özek (2015), Schwerdt et al. (2017), Figlio and Özek (2020) in Florida; Hwang and Koedel (2022) in Indiana; and Mumma and Winters (2023) in Mississippi.

or whose proficiency can be demonstrated with a teacher's portfolio. As such, nearly all existing RD studies on early grade retention rely on a fuzzy RD design to identify the effect of retention on student outcomes. Yet we do not know if these exemptions indeed identify students who are less likely to benefit from retention. We examine these research questions using student-level administrative data from Florida, which requires third graders to score at or above Level 2 (out of 5 achievement levels) on the statewide reading test to be promoted to fourth grade.

Our findings suggest that the benefits of retention (1) are larger for students with lower baseline reading achievement and (2) are indeed smaller for students exempt from retention. Together, these results imply that the average treatment effect on the treated (ATT) is much larger than the predicted effects that would come from removing the exemptions or increasing the passing threshold, i.e., the average treatment effect on the control (ATC). For example, we find that, as currently implemented, retaining students increases their sixth grade reading scores by 0.69σ, but further increasing the threshold by 50 points (0.8σ, roughly equivalent to moving the threshold from Level 2 to slightly above Level 3 on the third-grade reading test) and removing exemptions would have no impact on the sixth grade reading scores of the newly retained students. These findings also imply that existing studies on early grade retention policies that rely on traditional fuzzy RD designs significantly underestimate the benefits of retention. In particular, we show that the ATT estimates are roughly 20 percent larger than the LATE estimates of the retention effects on reading scores in grades 4 through 8.

II Model and Estimation Approach

II.A Underlying Model and Assumptions

We use as our base model one of the canonical models used to consider the effect of a binary treatment on a single outcome, and in particular the model that forms the basis for marginal treatment effect (MTE) estimation (e.g., Heckman (2010); Heckman and Vytlacil (2007a,b); Brinch et al. (2017); Mogstad et al. (2018); Kline and Walters (2019)). Specifically, we assume that each individual is defined by four variables: their outcome if they are not treated, the effect that the treatment has on their outcome, their propensity to enroll in the treatment, and their value of the

running variable; we denote these as μ_i, τ_i, η_i, and Z_i, respectively. In other words, we use μ_i to denote individual i's outcome in the absence of treatment and τ_i to denote the causal effect of the treatment on individual i's outcome; clearly $\mu_i + \tau_i$ is then their outcome if they are treated.

Letting T_i be a dummy variable denoting whether someone is in the treatment or control group, the observed outcome can be written as: $Y_i = \mu_i + \tau_i T_i$. As is common in the MTE literature, we further assume that treatment is determined according the following choice equation: $T_i = \mathbf{1}(\nu^*(Z_i) \geq \eta_i)$ for some (unknown) function of the running variable $\nu^*(Z_i)$. As a researcher, we observe Y_i, T_i, and Z_i, but do not observe the latent variables μ_i, τ_i, and η_i.

We then define following two conditional moments:

$$\mu^*(\eta, Z) = \mathbb{E}[\mu_i | \eta_i = \eta, Z_i = Z] \qquad \text{and} \qquad \tau^*(\eta, Z) = \mathbb{E}[\tau_i | \eta_i = \eta, Z_i = Z] \quad (1)$$

The function $\tau^*(\eta, Z)$, in particular, corresponds to the marginal treatment effect (MTE) function as defined in Heckman and Vytlacil (1999, 2005) and is generally the object of interest itself or, more commonly, the objects of interest can be derived from it. For example, full knowledge of the function $\tau^*(\eta, Z)$ would allow one to calculate the overall average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment effect on the compliers (LATE), and other estimands of interest. We use the star notation, i.e., denoting the functions as μ^* and τ^*, to distinguish the true conditional moment functions from generic potential conditional moment functions μ and τ.

While the conditional moment functions in Equation (1) correspond most closely with the objects of interest, they are a bit removed from what is observed in the data. We therefore also define two additional conditional moments, which are more closely related to what we observe. These moments are defined as follows:

$$k_0^*(\eta, Z) = \mathbb{E}[\mu_i | \eta_i > \eta, Z_i = Z] \qquad \text{and} \qquad k_1^*(\eta, Z) = \mathbb{E}[\mu_i + \tau_i | \eta_i \leq \eta, Z_i = Z] \quad (2)$$

In the method, we estimate $k = (k_0, k_1)$ and then transform this estimate into an estimate of τ. We denote the parameter space of k as $\mathcal{K} = \mathcal{K}_0 \times \mathcal{K}_1$ and endow each of the spaces $\mathcal{K}_i$ with the sup norm, i.e., $||k_i||_\infty = sup\{|k_i(\eta, Z)|\}$. Similarly, we will note the implied parameter space of τ as $\mathcal{T}$ and also endow it with the sup norm.

In the definitions above, we implicitly assume that the conditional first moments

exist. We make this assumption explicit below, along with the other assumptions we use:

Assumption 1. $\mathbb{E}[\mu_i^2|\eta_i, Z_i] < \infty$ *and* $\mathbb{E}[\tau_i^2|\eta_i, Z_i] < \infty$ *for all* $\eta_i \in (0,1)$ *and* $Z_i \in \mathbf{Z} \equiv (\underline{Z}, \overline{Z})$.

Assumption 2. Z_i *is continuously distributed over* $\mathbf{Z}$ *with a strictly positive distribution function and* ν_i *is continuously distributed conditional on* Z_i.

Assumption 3. *Both* $\mu^*(\eta_i, Z_i)$ *and* $\tau^*(\eta_i, Z_i)$ *are twice-continuously differentiable functions of* (η_i, Z_i).

Assumption 4. $\nu^*(Z_i) \in (0,1)$ *for all* Z_i *and there exists a* $Z^* \in \mathbf{Z}$ *such that:*

$$\lim_{Z_i \uparrow Z^*} \nu^*(Z_i) \equiv p_l < p_h \equiv \lim_{Z_i \downarrow Z^*} \nu^*(Z_i)$$

Assumption 5. *The parameter space* $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, *where:*[2]

$$\nu(Z_i) = \begin{cases} \nu_1(Z_i) & \text{if } Z_i < Z^* \\ \nu_2(Z_i) & \text{if } Z_i > Z^* \end{cases}$$

and where $\nu_i \in \mathcal{V}_i$. *Let* $\mathcal{W}_{2,2}$ *be the Sobolov space of functions* $f : \mathbf{Z} \to \mathbb{R}$ *and* $||f||_{\mathcal{W}_{2,2}}$ *be its norm, as defined in Freyberger and Masten (2019). Then* $\mathcal{V}_i = \{\nu_i \in \mathcal{W}_{2,2} : ||\nu_i||_{\mathcal{W}_{2,2}} \leq V_i\}$ *for some constant* V_i. *Furthermore, we will assume that* $\nu^* \in \mathcal{V}$.

Assumption 6. *The parameter space* $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$. *Let* $\mathcal{W}_{2,2}$ *be the Sobolov space of functions* $f : (0,1) \times \mathbf{Z} \to \mathbb{R}$ *and* $||f||_{\mathcal{W}_{2,2}}$ *be its norm, as defined in Freyberger and Masten (2019). Then* $\mathcal{K}_i = \{k_i \in \mathcal{W}_{2,2} : ||k_i||_{\mathcal{W}_{2,2}} \leq K_i\}$ *for some constant* K_i. *Furthermore, we will assume that* $k^* \in \mathcal{K}$.

Assumptions 1 and 2 are both relatively benign assumptions; the first ensures that we can use the standard asymptotic methods and the second ensures that asymptotically there will be a large number of observations arbitrarily close to each point $Z_i \in \mathbf{Z}$. The second also allows us to normalize η_i to be distributed uniformly from $(0,1)$. With this standard normalization, the cutoff value $\nu^*(Z_i)$ is equal to $Pr(T_i|Z_i)$, i.e. to the propensity score.

[2]Note here that we leave $\nu(Z^*)$ undefined.

The next two assumptions are reformulations of the standard assumptions required for RD designs. Assumption 3 captures the fact that the conditional moment functions are continuous around the discontinuity, although we extend the assumption such that the functions are continuous everywhere. We also require them to be twice-continuously differentiable, which ensures that our penalty functional described below to be well-defined. Assumption 4 captures the fact that it is a fuzzy RD context, in that there is a point Z^* at which the probability of treatment jumps discontinuously and that for every value of the running variable there are both treated and untreated individuals. We assume that there is a single discontinuity and that the probability increases as one moves across the threshold from left to right, but the first is easy to relax and the second is without loss of generality and so both are for expositional ease.

Assumptions 5 and 6 are that the parameter space is compact, which helps ensure that our non-parametric estimation approaches converge. As written, this assumptions permits us to consider uniform convergence and both could be relaxed if one was only interested in pointwise convergence. By assuming that the true functions, i.e., ν^* and k^*, fall within the parameter space, these also capture the assumption that the true functions are smooth.[3]

II.B Proposed Estimator

The key challenge is that the functions $k_i(\eta, Z)$ are not non-parametrically identified. Roughly speaking, we will obtain identification by greatly restricting the set of possible functions we consider. In particular, we will restrict the set of k functions to be those in which each of the $k_i(\eta, Z)$ functions are additively separable and linear in η, i.e., that $\mathcal{K}_R$ is defined as follows:[4]

$$\mathcal{K}_R = \{k \in \mathcal{K} | k_0(\eta, Z) = \alpha\eta + \gamma(Z) \;\&\; k_1(\eta, Z) = \beta\eta + \delta(Z)\} \tag{3}$$

We discuss in the next section ways to this restriction can be justified and the estimator interpreted even if the true functions do not satisfy this restriction, i.e., even

[3]This makes Assumption 3 redundant, but we included to more explicitly capture one of the main assumptions of an RD design.

[4]This is equivalent to the restriction that $\mu(\eta, Z) = \gamma(Z) + \alpha(1 - 2\eta)$ and $\tau(\eta, Z) = \delta(Z) - \gamma(Z) + \alpha + 2\eta(\beta - \alpha)$.

if $k^* \notin \mathcal{K}_R$.

Given this restriction, we define our estimator as follows:

Global Regression Discontinuity Design. *Our proposed estimator consists of three steps:*

1. *Estimate $\nu(Z_i)$ as follows:*

$$\hat{\nu} = \arg\min_{\nu \in \mathcal{V}} \left\{ \sum_{\forall i} \Big(T_i - \mathbf{1}(Z_i > Z^*) \cdot \nu(Z_i) \big) - \mathbf{1}(Z_i < Z^*) \cdot \nu(Z_i) \Big)^2 + J_\nu(\nu) \right\} \quad (4)$$

   for some penalty functional J_ν.

2. *Estimate k as follows:*

$$\hat{k} = \arg\min_{k \in \mathcal{K}_R} \left\{ \sum_{\forall i} \Big(Y_i - (1-T_i) \cdot k_0\big(\hat{\nu}(Z_i), Z_i\big) - T_i \cdot k_1\big(\hat{\nu}(Z_i), Z_i\big) \Big)^2 + J_k(k) \right\} \quad (5)$$

   for some penalty functional J_k.

3. *Estimate τ using $\hat{k}$ as follows:*

$$\hat{\tau}(\eta, Z) = T\big(\hat{k}_0, \hat{k}_1\big) \quad (6)$$

   where:

$$T(k_0, k_1) = k_1 - k_0 + \eta \frac{\partial k_1}{\partial \eta} + (1-\eta) \frac{\partial k_0}{\partial \eta} \quad (7)$$

$$= \delta(Z) - \gamma(Z) + 2(\beta - \alpha)\eta + \alpha \quad (8)$$

In our implementation, we let the the penalty functionals take the form:

$$J_\nu(\nu) = \int \Big[\nu''(Z) \Big]^2 dZ \quad (9)$$

$$J_k(k) = \int \Big[\gamma''(Z) \Big]^2 dZ + \int \Big[\delta''(Z) - \gamma''(Z) \Big]^2 dZ \quad (10)$$

Most of the theoretical results are not dependent on this particular form of the penalty function, however, and generally apply to any non-parametric approach to estimate $\hat{\nu}$, $\hat{\delta}$, and $\hat{\gamma}$. Note also that we penalize $\int [\delta''(Z) - \gamma''(Z)]^2 dZ$ rather than $\int [\delta''(Z)]^2 dZ$

directly; this could roughly be thought of as separately penalizing the functions μ and τ as defined in Equation (1) rather than separately penalizing the functions k_0 and k_1 as defined in Equation (2). Again, this choice matters for the implementation, but not for the theoretical results.

Before discussing the interpretation of the estimator in the next section, we start by showing that it is well-defined and consistent. That is, with enough data – generated according the assumptions in Section II.A – the estimator defined above will result in an estimate of the marginal treatment effect (MTE) function that becomes arbitrarily close to a unique MTE function, which we denote τ^*_{GRDD}.[5] The formal theorem and a sketch of the proof are below, with the full proof in Appendix A:

Proposition 1. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined above. Then given Assumptions 1 - 6, there exists τ^*_{GRDD} such that $\hat{\tau}_{GRDD} \xrightarrow{p} \tau^*_{GRDD}$.*

Proof Sketch. Our approach is to first show that there exists $k^*_{GRDD} \in \mathcal{K}_R$ such that that $\hat{k} \xrightarrow{p} k^*_{GRDD}$ and then appeal to the fact that T is continuous to conclude that $\hat{\tau}_{GRDD} = T(\hat{k}) \xrightarrow{p} T(k^*) \equiv \tau^*_{GRDD}$.

Showing that there exists such a k^*_{GRDD} can be done using the basic results about extremum estimators, e.g., Theorem 2.1 of Newey and McFadden (1994). The only component unique to this context is to show that with the restriction that $k \in \mathcal{K}_R$ we obtain identification, i.e., that there is a unique minimizer of the function that Equation (5) converges to. For this, we rely on the fact that the probability of treatment jumps at the discontinuity, which pins down the linear terms α and β. Given these linear terms, we then can choose $\gamma(Z)$ and $\delta(Z)$ to match the rest of the observed moments and therefore minimize the function. □

Based on this result, we will ignore the statistical uncertainty in the next section and focus on ways to interpret and motivate τ^*_{GRDD}. We note, however, that the method can improve precision over a traditional RDD; see Mulhern et al. (2023) for more discussion about the relative precision of the estimates.

[5] To be clear, this choice of norms described in Section II.A implies that we define "arbitrarily close" in a uniform sense, rather than a pointwise sense. Specifically, $\hat{\tau} \xrightarrow{p} \hat{\tau}^*$ means that for all $\epsilon, \delta > 0$ there exists an $\overline{N}$ such that $\forall n \geq \overline{N}$ we have that $\mathbb{P}\big(sup\{|\hat{\tau}_n(\eta, Z) - \hat{\tau}^*(\eta, Z)| : (\eta, Z) \in (0,1) \times \mathbf{Z}\} > \epsilon\big) < \delta$.

III Interpreting and Motivating the Estimator

In the previous section, we introduced the estimator and showed that it converges. The fact it converges to a MTE function does not imply that it converges to the true MTE function; however, and so we now discuss various ways this estimator can be motivated and interpreted. First, we highlight that locally the estimator converges to the true treatment effect even though it assumes – potentially incorrectly – that the conditional moment functions are additively separable and linear in η. Formally, we get that following theorem:[6]

Proposition 2. *The estimated effect on the set of compliers at the Z^* converges to the true effect on that set, i.e.:*

$$\frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*)d\eta = \frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*(\eta, Z^*)d\eta \tag{11}$$

Of course, if all one was concerned about was the local average treatment effect, one could instead use a traditional fuzzy RD estimator. In contrast to a traditional fuzzy RD design, however, the global RD design also provides effect estimates away from the complier population at the cutoff. We first discuss two alternative ways in which these estimates can be interpreted and then discuss three ways our proposed estimator can be motivated.

To do so, we first define two natural alternatives: (1) an observational study in which one simply compares the the treatment average to control average at every point $Z \in \mathbf{Z}$ and (2) a traditional fuzzy regression discontinuity design. Formally, we get that:

$$\tau^*_{obs}(\eta, Z) = \mathbb{E}[Y_i|T_i = 1, Z_i = Z] - \mathbb{E}[Y_i|T_i = 0, Z_i = Z] \tag{12}$$

$$\tau^*_{RDD}(\eta, Z) = \frac{1}{p_h - p_l}\left(\lim_{Z\downarrow Z^*} \mathbb{E}[Y_i|Z_i = Z] - \lim_{Z\uparrow Z^*} \mathbb{E}[Y_i|Z_i = Z]\right) \tag{13}$$

Note that although we define τ^*_{obs} and τ^*_{RDD} to be functions of (η, Z), τ^*_{obs} will not vary based on η and τ^*_{RDD} will not vary based on η or Z.

In what follows, we will generally focus on the conditional average treatment effect

[6]This result is in many ways the RD version of Theorem 1 of Kline and Walters (2019); however, since the RD design requires using observations "near" Z^* to infer the moments at Z, the result only holds asymptotically.

(CATE) – conditioning on the value of the running variable – or $\tau(Z) \equiv \int_0^1 \tau(\eta, Z)d\eta$. The general idea can be extended to other estimands of interest and, given the conditional effect estimates, it is straightforward to aggregate them to estimates of the overall average treatment effect (ATE) or overall treatment on the treatment (ATT). The interpretations and motivations discussed below therefore also hold for the overall ATE, ATT, and other estimands of interest.

III.A Interpreting the Estimator

We start by discussing two potential alternative estimators and show that they are in fact equivalent to the one proposed in Subsection II.B.

Bias-Adjusted Observational Study

Consider an alternative approach, where instead of employing the estimator defined in Section II.B we start with the traditional estimator that simply uses the observational data, i.e., $\tau^*_{obs}(Z)$. We then start by noting that this estimate is – by definition – the true CATE plus a bias term. Given an estimate of the bias term, it would therefore make sense to estimate the CATE as $\tau^*_{obs}(Z)$ minus an estimate of the bias term. Of course, estimating the bias is quite challenging, but here we have information at the discontinuity. In particular, we observe the true average treatment effect on the compliers at the discontinuity. Roughly speaking, it seems reasonable to estimate the bias in $\tau^*_{obs}(Z)$ by comparing the fuzzy RD estimate of the LATE to the $\tau^*_{obs}(Z)$ estimates at the discontinuity. If the fuzzy RD estimates are identical to the observational estimates at the cutoff, this suggests that the observational estimates have minimal bias. In contrast, if the fuzzy RD design diverges from them at the cutoff, this suggests that the observational estimates estimates are quite biased.

The arguments above suggests an approach where one: (i) first ignores the discontinuity and uses the observational data to generate an estimate of the CATE function using traditional approaches; (ii) then estimates the LATE using a traditional fuzzy RD approach; (iii) then estimates the bias in the initially estimated CATE function by comparing the LATE estimates to the CATE estimates at the discontinuity; and (iv) generates the final estimates by adjusting the CATE estimates for the estimated bias. While this seems a very different approach than the one defined in Section II.B, as shown in the theorem below one can think of the proposed estimator as doing exactly that.

Remark 1. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and $\tau^*_{obs}(Z)$ be the estimate generated from the traditional observational study, as defined in Equation (12). We then have:*

$$\tau^*_{GRDD}(Z) = \tau^*_{obs}(Z) - b \tag{14}$$

where b is a measure of the bias in the observational estimates. Specifically, we have:

$$b = \xi_h \cdot \left(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\right) + \xi_l \cdot \left(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\right) \tag{15}$$

where $\xi_i \in \mathbb{R}$ is a function of p_h, p_l and $\nu(Z)$, and $\tau^*_{obs}(Z^*_h) = \lim_{Z \downarrow Z^*} \tau^*_{obs}(Z)$ and $\tau^*_{obs}(Z^*_l) = \lim_{Z \uparrow Z^*} \tau^*_{obs}(Z)$.

Extrapolated Regression Discontinuity Design

Next, consider a third approach where one starts with the LATE estimate generated by the fuzzy regression discontinuity design. As discussed, however, the LATE estimate is generally not sufficient because it is the local average treatment effect rather than the global average; in the fuzzy RD context, locality refers both to the estimate being local to the set of compliers and to the estimate being local to the discontinuity. A natural approach is therefore to adjust the LATE estimate by first extrapolating to an estimate of the CATE at the discontinuity and then extrapolating away from the discontinuity.

Again, the challenge here is clearly how one can extrapolate from the LATE to an estimate of the CATE at the discontinuity, as well as how one can extrapolate the CATE away from the discontinuity. To guide this, however, there has been some recent work exploring how to extrapolate from the average treatment effect on the compliers to the average treatment effect on everyone (Brinch et al., 2017; Mogstad et al., 2018; Kowalski, forthcoming; Opper, 2023). Furthermore, it seems reasonable to extrapolate away from the discontinuity by comparing the $\tau^*_{obs}(Z)$ estimates. If we consider the simple case in which the propensity scores are constant away from the discontinuity, it seems reasonable to infer that if $\tau^*_{obs}(Z)$ is larger than than $\tau^*_{obs}(Z')$ then it is more likely than not that the CATE at Z is larger than the CATE at Z'.

In short, the above suggests an estimation approach that: (i) first estimates the LATE using a traditional fuzzy RD approach; (ii) using the conditional moments at the discontinuity, extrapolate the LATE estimate to an estimate of CATE using the

linear approach in Brinch et al. (2017) and Kowalski (forthcoming); (iii) generate an estimate of the $\tau^*_{obs}(Z)$ function using traditional approaches; and finally (iv) use the estimates of the $\tau^*_{obs}(Z)$ function to adjust for differences in the CATE at the discontinuity to those away from the discontinuity. Again, at first glance this seems a very different approach than the one defined in Section II.B, but as we show in the theorem below one could also think of the proposed estimator as doing exactly that.

Remark 2. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and τ^*_{RDD} be the estimate generated from the traditional regression discontinuity design, as defined in Equation (13). We then have:*

$$\tau^*_{GRDD}(Z) = \tau^*_{RDD} + l_C + l_Z \tag{16}$$

where l_C adjusts for fact that τ^*_{RDD} is local to the set of compliers and l_Z adjusts for the fact that τ^*_{RDD} is local to the discontinuity. Specifically, we have that:

$$l_C = \left[\beta^* \cdot (1 - p_h) + k^*_1(p_h, Z^*)\right] - \left[k^*_0(p_l, Z^*) - \alpha^* \cdot p_l\right] - \tau^*_{RDD} \tag{17}$$

$$l_Z = \tau^*_{obs}(Z) - \tau^*_{obs}(Z^*_h) + (\alpha^* - \beta^*) \cdot \Big(\nu(Z) - p_h\Big) \tag{18}$$

where $\alpha^* = \frac{k^*_0(p_h,Z^*) - k^*_0(p_l,Z^*)}{p_h - p_l}$, $\beta^* = \frac{k^*_1(p_h,Z^*) - k^*_1(p_l,Z^*)}{p_h - p_l}$.

III.B Motivating the Estimator

While the above subsection suggests that the proposed estimator can be interpreted in two alternative and generally intuitive ways, it does not provide explicit rationale for the proposed estimator. We now discuss three such motivations.

Additional Strong Assumptions

For this motivation, we simply add the required assumptions need to ensure that the estimated MTE function our estimator converges to is in fact the true MTE function. This result can be stated succinctly in the following theorem:

Proposition 3. *Suppose that $k^* \in \mathcal{K}_R$. Then the estimated MTE function converges to the true MTE function, i.e., $\tau^*_{GRDD} = \tau^*$.*

The assumptions required for Theorem 3 are, in our opinion, relatively strong. Assuming that $k^* \in \mathcal{K}_R$ amount to assuming that the true conditional moment functions are indeed additively separable and linear in ν. It is worth emphasizing, however,

that the assumption becomes much more tenable with the addition of covariates. For example, if we only aim to extrapolate away from the cutoff and not to the set of non-compliers, the assumption needed to generate consistent estimates away from the cutoff is weaker than in Angrist and Rokkanen (2015). It is not the only way to motivate the estimator, however, and we turn next to alternative motivations.

Preference for Less Complex Models

One motivation for the proposed estimator stems from a general preference for less complex models (e.g., Hastie et al. (2009)). To formalize this, we need to develop a way to compare the relative complexity of two models; in other words, we need to define a relation $\preceq$ such that $k \preceq k'$ iff k is a less complex model than k'. We will do so, by first defining a triple for each model – $c(k) = (c_{11}, c_{12}, c_{22})$ – where:

$$c_{11} = \mathbb{E}\left[\sum_{i=0,1}\left(\frac{\partial^2}{\partial\eta^2}k_i(\eta, Z)\right)^2\right] \quad c_{12} = \mathbb{E}\left[\sum_{i=0,1}\left(\frac{\partial^2}{\partial\eta\partial Z}k_i(\eta, Z)\right)^2\right] \quad c_{22} = \mathbb{E}\left[\sum_{i=0,1}\left(\frac{\partial^2}{\partial Z^2}k_i(\eta, Z)\right)^2\right]$$

We then use a lexicographic order to compare the complexity of any two models. Specifically, we say that model k is a less complex model than k' – i.e., $k \preceq k'$ – if: (i) $c_{11}(k) < c_{11}(k')$; (ii) $c_{11}(k) = c_{11}(k')$ and $c_{12}(k) < c_{12}(k')$; or (iii) $c_{11}(k) = c_{11}(k')$ and $c_{12}(k) = c_{12}(k')$ and $c_{22}(k) \leq c_{22}(k')$. It follows that we can similarly write that k is a strictly less complex model than k' – i.e., $k \prec k'$ – if $k \preceq k'$ and $c_{22}(k) < c_{22}(k')$.

Given this ordering, we can then motivate our estimator as being the least complex model that is consistent with the observed data. More formally, we have the following theorem:

Proposition 4. *Define $\mathcal{K}_D$ to be the k functions that are consistent with the true conditional moments, i.e.,*

$$\mathcal{K}_D = \left\{k : \mathbb{E}[Y_i|\nu(Z_i), Z_i, T_i] = (1-T_i)\cdot k_0\big(\nu(Z_i), Z_i\big) + T_i\cdot k_1\big(\nu(Z_i), Z_i\big) \text{ for all } z_i \in \mathbf{Z}\right\}$$

Then $\tau^*_{GRDD} = T(k^*_{GRDD})$ where k^*_{GRDD} is the minimum element of the preordered set $\big(\mathcal{K}_D, \preceq\big)$ when the preorder $\preceq$ is defined as the lexicographic order on $c(k)$, defined above.

While we leave the proof of the theorem to Appendix A, we highlight a few aspects of the theorem here. First, note that the proof states that k^*_{GRDD} is the minimum element rather than a minimal element: In other words, the theorem states that every

other model k consistent with the observed data is *strictly* more complex than k^*_{GRDD}.

We also want to acknowledge explicitly that it is not obvious that the lexicographic ordering is the correct way to measure model complexity nor are we even arguing that it is the right way to do so. Instead, the theorem states that *if* one believes that it is the correct way to measure model complexity, then the estimator proposed in Section II.B gives the least complex model consistent with the data. If one wants to prioritize this motivation (the least complex model consistent with the observed data) but prefers a different measure of model complexity, e.g., where the complexity of model k is measured as $c_{11}(k) + c_{12}(k) + c_{22}(k)$, one could adjust the estimator defined in Section II.B to allow for alternative measures of model complexity; see Appendix B for the details. We have opted to define model complexity using the lexicograph ordering as it makes it more clear how the conditional moments translate into estimated effects – again, see Appendix B for more discussion – and because it more readily permits the other motivations discussed below.[7]

Preferable to the Alternatives

As a final motivation, we compare the proposed estimator to alternative approaches – with a particular focus on the purely observational study and the traditional fuzzy regression discontinuity design defined formally in Equations (12) and (13). A challenge is that we cannot expect the proposed estimator to be preferable under all data generating processes; there are simply too many plausible data generating processes that are consistent with what is observed by the researcher for any one estimator to dominate all others under all data generating processes.[8] We therefore have to limit ourselves to the hope that the proposed estimator is preferable to the alternatives *on average* or under most data generating process. Of course, doing so requires that we define a probability measure over the data generating processes itself – as opposed to just over the observed data conditional on the true underlying moments – and a more precise definition of "preferable."

[7]While we specify the lexicographic ordering such that we first compare $c_{11}(k)$ to $c_{11}(k')$ and then compare $c_{12}(k)$ to $c_{12}(k')$, this decision is not particularly important and the proof under the canonical fuzzy RDD setup also holds if the ordering under consideration instead first compares $c_{12}(k)$ to $c_{12}(k')$ and then compares $c_{11}(k)$ to $c_{11}(k')$. However, that ordering does become important if there are instead multiple discontinuities.

[8]For example, a traditional fuzzy RD design is preferable to the proposed estimator when the true model is one in which the treatment effect is identical for everyone and selection varies based on the value of the running variable; similarly, a purely observational study is preferable when the true model is one in which there is large endogenous selection into the treatment at the discontinuity and little selection away from the discontinuity.

To do so, we will assume that the conditional moments are generated according to a modified Gaussian process (GP). Specifically, let $k(\eta_i, Z_i, T_i) = T_i k_1(\eta_i, Z_i) + (1 - T_i) \cdot k_0(\eta_i, Z_i)$ and $h(\eta_i, Z_i)$ be the 1×2 vector $[1, \eta_i]$. We then assume that:

$$k_j(\eta, Z) = h(\eta, Z) \cdot \beta_j + \tilde{k}_j(\eta, Z) \tag{19}$$

where $\tilde{k}_j(\eta, Z)$ is generated according to a Gaussian process with mean zero and covariance C_j for $j \in \{0, 1\}$.[9] For the comparison with the traditional RD design, we will restrict the class of C_j functions by assuming that C_j is stationary in Z (but not necessarily in η) and that there is a large enough direct effect of Z on the moments (as opposed to the interaction between Z and η). Formally, we make the assumption that:

$$C_j\big((Z, \eta), (Z', \eta')\big) = C_{j,Z}(|Z - Z'|) C_{j,\eta}(\eta, \eta') \tag{20}$$

with $C_{j,Z}(|Z - Z'|) > 0$ decreasing in $|Z - Z'|$ and $C_{j,\eta}(\eta, \eta') = c_\eta + f(\eta, \eta') + > 0$ with f decreasing in η and increasing in η' if $\eta > \eta'$ and where c_η – which corresponds to the direct effect – is sufficiently large relative to the interaction term – i.e., $\max f(\eta, \eta')$. Finally, we will assume that $\beta_j \sim N(0, \Sigma_\beta)$ and will consider the uninformative prior limit in which $\Sigma_\beta^{-1} \to \mathbf{0}$ where $\mathbf{0}$ is the zero-matrix.

In contrast to k, we will consider the function $\nu(Z)$ to be fixed and further assume that (conditional on k) the rest of the data generating process accords to the assumptions outlined in Section II.A.[10] As before, we can transform any realization of k into the implied marginal treatment effect function according to the bounded linear operator T, defined in II.B, and so this process also defines a data generating process over the marginal treatment effect functions.

As in Mogstad et al. (2018), we assume that the researcher is interested in some summary measure of $\hat{\tau}^*$, defined as:

$$\Gamma(\tau^*) = \int_{\mathbf{Z}} \int_0^1 \tau^*(\eta, Z) \omega(\eta, Z) d\eta dZ \tag{21}$$

[9] The covariance is often referred as the "kernel" and denoted as K. We use the term "covariance" and denote it as C to ensure it does not get confused with the conditional moment functions k_i.

[10] We acknowledge some tension between the modified GP which generates k and the assumptions in Section II.A that the conditional moments are bounded and twice-continuously differentiable. We will ignore these subtleties for now, e.g., by implicitly putting restrictions on C to ensure the sample paths are twice-continuously differentiable and allowing the bounds to differ for each realization of the modified GP.

for some weighting scheme $\omega(\eta, Z)$. For example, if $\omega(\eta, Z)$ equals the pdf of Z_i, then $\Gamma(\tau^*)$ corresponds to the ATE. As discussed in Mogstad et al. (2018), different weighting schemes generate other such parameters of interest. Given a realization of the modified GP and its implied MTE function τ^* as well as an estimate of the MTE function $\hat{\tau}^*$, we then define the loss as:

$$l_k(\hat{\tau}^*, \tau^*) = \Big(\Gamma(\tau^*) - \Gamma(\hat{\tau}^*)\Big)^2 \tag{22}$$

We subscript the loss by k to make clear that $l_k(\hat{\tau}^*, \tau^*)$ depends on the realization of the modified GP. It is natural, therefore, to evaluate the performance of an estimator using the expected loss, where the expectation is taken over realization of the modified GP. Formally, consider any estimator $\hat{\tau}_{alg}$, i.e., an algorithm to transform the observed data to an estimate of the MTE function, and define its expected loss as:

$$\mathscr{L}_C(\hat{\tau}_{alg}) = \mathbb{E}[l_k(\hat{\tau}^*, \tau^*)] \tag{23}$$

Note this depends on the covariance matrix of the GP, i.e., on C, since the expectation depends on the assumed hyperparameters of the GP. We can now precisely define the way in which the global RD design is preferred to the alternative approaches, as evidence by the following theorem:

Proposition 5. *We say that the estimator $\hat{\tau}_a$ **dominates** $\hat{\tau}_b$ iff $\mathscr{L}_C(\hat{\tau}_a) < \mathscr{L}_C(\hat{\tau}_b)$ for all C. Then:*

1. *The global RD design dominates the traditional observational study.*
2. *If $\nu(Z) = p_h$ above the cutoff and $\nu(Z) = p_l$ below the cutoff, the global RD design dominates the traditional fuzzy regression discontinuity design.*
3. *No estimator dominates the global RD design.*

Proof Sketch. For the first statement – that the global RD design dominates the traditional observational study – we use the fact that the only difference between the two estimates is a bias measure b; see Theorem 1. We then show that conditional on the four moments observed at the discontinuity – i.e., $k_0(p_l, Z^*)$, $k_0(p_h, Z^*)$, $k_1(p_l, Z^*)$, and $k_1(p_h, Z^*)$ – the global RDD approach generates unbiased estimates of $\tau(Z)$. This suggests that the only difference between the two estimates is a bias term that exists in

the observational study and is absent in the global RD design, which implies that the global RDD estimates are preferred. Since this is true regardless of which moments at the discontinuity are observed, it is true in expectation regardless of C.

For the second statement – that the global RD design dominates the traditional fuzzy regression discontinuity design – we start by showing that instead of comparing the global RDD estimator to a traditional RDD estimator, we can compare the global RDD estimator to one that extrapolates from the observed moments at the cutoff to the overall CATE at the cutoff, but – unlike the global RDD – does not use information away from the cutoff to extrapolate away from the cutoff. Ignoring this information means that expected loss is large as long as there is the possibility of direct effects of the running variable on the condition moments. In this case, the proposed global RDD dominates the version in which information away from the threshold is ignored, which in turn dominates the traditional RDD. In other words, the condition that the direct effect is large enough along with the assumption that $\nu(Z) = p_h$ above the cutoff and $\nu(Z) = p_l$ below the cutoff, ensures that the best information about $k_1(1, Z)$ – which we do not observe – is $k_1(\nu(Z), Z)$ rather than some other value $k_1(\nu(Z'), Z')$.

For the final statement – that no estimator dominates the global RD design – we use the fact that if the true effects functions are separable and linear in η then the global RDD gives rise to the true effect function; see Theorem 3. In this case, it is clear that no estimator would be better and if $C = C_z$ then the true effect functions are guaranteed to be separable and linear in η. Thus, if $C = C_z$ then the global RDD estimator minimizes $\mathscr{L}_C(\hat{\tau}_{alg})$ and so no estimator can dominate it.

□

IV Empirical Application: Grade Retention Policies

In this section, we present an application of our estimator in education policy: a field where fuzzy RD design has become more popular with the increasing use of student test score cutoffs (or performance index cutoffs based on student test scores) to identify eligibility for educational interventions. In particular, we explore the broader effects of test-based retention policies. As we detail below, there is extensive

literature examining the effects of grade retention on student outcomes using fuzzy RD designs; however, these estimated effects often apply only to compliers (i.e., students not exempt from retention) right below retention cutoffs. In this exercise, we ask whether these effects differ for exempt students and for lower-performing students identified for retention.

IV.A Policy Background and Data

Calls to end social promotion in schools in the 1990s and an increased popularity of educational accountability and standardized testing led to test-based retention policies in many states and school districts in the United States over the past three decades. Perhaps the most influential of these policies has been Florida's third grade retention policy, which was enacted in 2002 and provided the blueprint for others nationwide. This policy requires students who score in the lowest achievement level on statewide reading test to repeat third grade and receive instructional support (e.g., additional instruction time in reading, being assigned to highly effective teachers).

There are several "good cause exemptions" that allow students to be promoted to the fourth grade despite failing to score at the Level 2 benchmark or above. In particular, students in the lowest achievement level in reading can be promoted to fourth grade (1) if they have been in the English learner program for less than two years; (2) if they have certain disabilities and have been already retained once until third grade; (3) if they have received intensive reading remediation for two years and have already been retained twice between kindergarten and third grade; (4) if they demonstrate that they are reading at a level equal to or above a Level 2 on the statewide reading test by performing at an acceptable level on an alternative standardized reading assessment approved by the State Board of Education; or (5) if they demonstrate proficiency through a teacher-developed portfolio. Despite these exemptions, the policy has affected a significant share of third graders in the state: in the first year of the policy, 21 percent of third graders were flagged for retention (i.e., scored below the retention cutoff) and 15 percent had to repeat third grade (Licalsi et al. (2019)). Among those flagged for retention, one-third received an exemption and were promoted to fourth grade. While retention rates gradually declined partly due to improvements in reading achievement and the increase in exemption rates, they remained sizable with roughly 10 percent of the third graders being retained in

2021-22 school year.

Several studies have examined the effects of being retained (and receiving instructional support) under Florida's retention policy on student outcomes using the discontinuity in retention likelihood and RD designs (Greene and Winters (2007), Winters and Greene (2012), Özek (2015), Schwerdt et al. (2017), Figlio and Özek (2020)). The overarching conclusion is that retained students outperform their same-age peers in the short term (one to three years), these achievement gains fade out over time. That said, retained students under Florida's retention policy significantly outperform their promoted peers when they reach the same grade level, and are also less likely to be retained in a later grade. While providing compelling evidence, by using traditional RD designs these papers all focus on the complier population at the discontinuity. In this paper, we use the proposed estimator to determine how these benefits differ for students away from the cutoff and for students who were promoted to fourth grade using exemptions.

To address these questions, we use student-level administrative data from a large urban school district (LUSD) in Florida. In our analysis, we use students who entered third grade for the first time between 2005-06 and 2010-11 school years and follow them until 8th grade. Roughly 17 percent of these students were flagged for retention and of those identified for retention, 38 percent were retained, corresponding to 7 percent of the third graders in these cohorts. Of those who were not flagged for retention, a small number of students (1 percent) were retained regardless and so there is two-sided non-compliance in this setting. Our main outcomes of interest are standardized reading scores in grades 4 through 8.[11]

IV.B Results

Figures 1 and 2 (along with Table 1) present the estimated effects for exempt and non-exempt students around and away from the retention cutoff in different ways. The overarching conclusion from this analysis is that the impact of retaining students

[11]In the analysis that follows, we use a same-grade comparison: That is, we compare the test scores of retained and promoted students when they reach the same grade level. Another approach commonly used in the grade retention literature is to compare the test scores of treated and comparison students in years following the treatment (i.e., same-age comparison). We prefer the former approach as we see additional time provided to retained students as part of the treatment. That said, we also conducted a same-age comparison (results available upon request) and the main conclusions remain unchanged.

is largest for those with the lower third grade reading scores and for those who – conditional on their third-grade reading score – are most likely to be retained.

For example, Figure 1 shows how the conditional average treatment effect on the treated individuals depends on third grade reading scores. Specifically, the solid line shows how the estimated effect – formally $\mathbb{E}[\tau_i|Z_i = Z, T_i = 1]$) – depends on the value of the running variable (Z_i, third grade reading scale scores centered at the retention cutoff). The dashed lines indicate the 95 percent point-wise confidence interval and were using a Bayesian bootstrap procedure, in which we repeatedly (n = 100) drew weights for each student from a Dirichlet distribution and estimate $\hat{\tau}$ using the procedure defined in Section II.B. The dashed lines then illustrate the range of these estimates.[12]

The results in Figure 1 suggest that the positive effects of retention on fourth grade reading scores monotonically decline with students' baseline reading achievement. At the cutoff, we find that retention increases fourth grade reading scores by roughly 0.9σ, which is consistent with the effect sizes found in the previous literature (Schwerdt et al. (2017), Figlio and Özek (2020)). This benefit grows to 1.2σ for students whose third grade reading scores fell 25 points below the cutoff, and to 1.4σ for students 50 points below the cutoff. In contrast, the positive effects decline to 0.8σ for students 25 points above the cutoff and to 0.6σ for those 50 points above. Since most students who are retained are below the cutoff, these findings suggest that the LATE estimates presented in prior RD studies in this context significantly underestimate the overall benefits of retention in the short term.

It is also clear from Figure 1 that at the discontinuity, the effect on the treated individuals jumps. This stems from the fact that, by construction, the characteristics of the treated population discontinuously change at the threshold. We illustrate the effect of this more directly in Figure 2, which illustrates how $\hat{\tau}(\eta, Z)$ varies by both Z and η. In this exercise, η_i can be interpreted as "promotion likelihood": a student is retained if and only if their η_i falls below a given cutoff. In other words, effect estimates for higher values of η_i indicate the retention effect for students who are least likely to be retained and vice versa. In this graph, each line corresponds to a set of (η_i, Z_i) values with the same estimated effect. There are two important takeaways from this figure.

[12]We used a Bayesian bootstrap instead of a traditional bootstrap to ensure that in every iteration there was two-sided imperfect compliance at the discontinuity.

First, consistent with Figure 1, the estimated effect declines as students' baseline reading achievement increases (moving from left to right). Second, we also observe that students who are less likely to receive an exemption and be promoted to fourth grade benefit significantly more from retention. For example, at the retention cutoff, the average effect for students who are least likely to be retained (η_i=1) is roughly 0.3σ where the average effect for those most likely to be retained (η_i=0) is 1σ. This finding suggest that the exemptions to the retention rule incorporated into Florida's policy indeed identify students who are least likely to benefit from retention. That said, exempt students with lower baseline achievement would still benefit from retention: the effect of retention on students 25 points below the cutoff who are most likely to receive an exemption is nearly 0.7σ while the effect for exempt students 50 points below the cutoff is roughly equivalent to the effect of retention for non-exempt students at the cutoff.

Table 1 extends this analysis to reading scores in grades 4 through 8 under different scenarios for treatment assignment. In the first column, we present the treatment effects under optimal treatment assignment (i.e., keeping the retention rate constant, yet assigning the individuals who would benefit most from retention). The second column presents the average treatment effect under the realized treatment assignment (average treatment effect on the treated or ATT); the third column gives the estimated effect on the complier population at the threshold (or LATE); the fourth gives the average treatment effect if students were randomly retained (overall average treatment effect on the treated or ATE); and the last column provides the average treatment effect if those who are not retained under the realized treatment assignment were retained (average treatment effect on the controls or ATC).

The results suggest that the realized assignment is nearly equivalent to the optimal assignment. In particular, average treatment effects under realized assignment are larger than 77 percent of the average treatment effects under optimal assignment in all cases. It also suggests that the ATT is larger than the LATE, implying that the policy increases test scores of those retained by more than has been shown in previous studies which use an RD approach to identify the LATE. However, the results in the last column shows that expanding the program would have minimal effects in the years after the student was retained and that these effects fade-out completely by sixth-grade. This suggests that Florida's policy is quite successful in identifying students most likely to benefit from retention.

V Conclusion

The trade-off between internal and external validity is a common issue in causal inference. In the context of RD design, this trade-off manifests itself in two ways. First, the RD estimates obtained using traditional methods only apply to individuals identified for treatment within a small bandwidth around the treatment cutoff. Second, in many RD applications, treatment assignment is fuzzy: that is, being on the treatment side of the cutoff does not fully determine treatment status due to non-compliance or policy-dictated exemptions. In those settings, it is hard to generalize traditional RD estimates to non-compliers. That said, understanding treatment effects beyond compliers around the treatment cutoff is critical from a public policy perspective in many settings for several reasons.

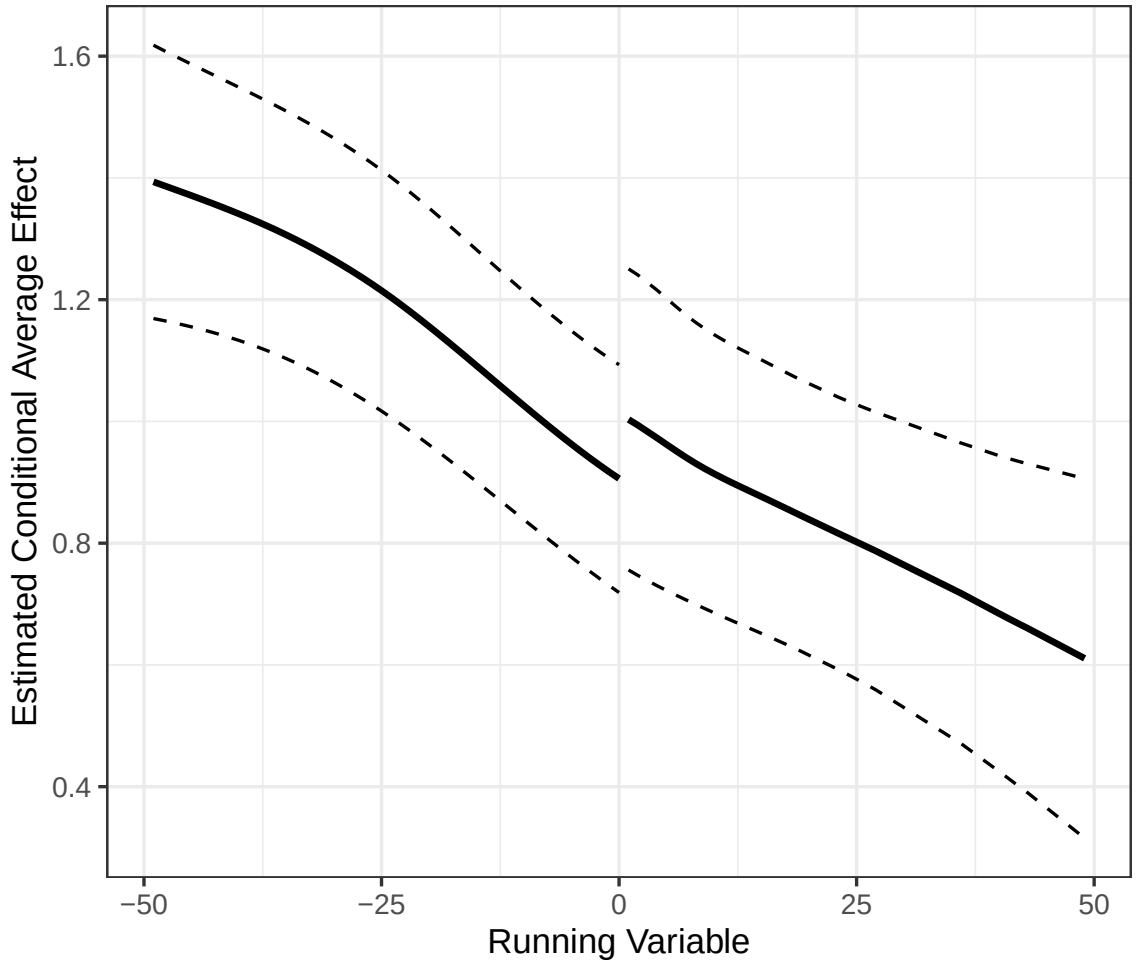
In this study, we propose a new method for use in fuzzy RD settings, which we call global regression discontinuity design, to address this issue. The estimator can be thought of either as a bias-adjusted observational study or an extrapolation of the traditional fuzzy regression discontinuity estimate (first to non-compliers at the cutoff and then to individuals away from the cutoff). We show that it can be motivated as being the least complex model consistent with the data or as an estimator that generates better estimates (on average) than either of the traditional approaches. We further show theoretically that no other estimators consistently generate better estimates than our proposed estimator.

We then present an application of this method in education policy. In particular, we examine the broader effects of early grade retention policies, which often require students to score above a predetermined threshold on third-grade reading tests to be promoted to fourth grade, on student outcomes using student-level data from Florida. Several prior students have addressed this question using traditional RD designs and found significant benefits. Here, we ask how these benefits differ for lower-performing students away from the cutoff and for low-performing students who were promoted using exemptions. We find that the positive effects of retention are larger for students with lower baseline reading achievement and smaller for student exempt from retention. Our findings also suggest that retaining more students, by either increasing the threshold or removing exemptions, would have limited effect on the newly retained students.

VI Graphs and Tables

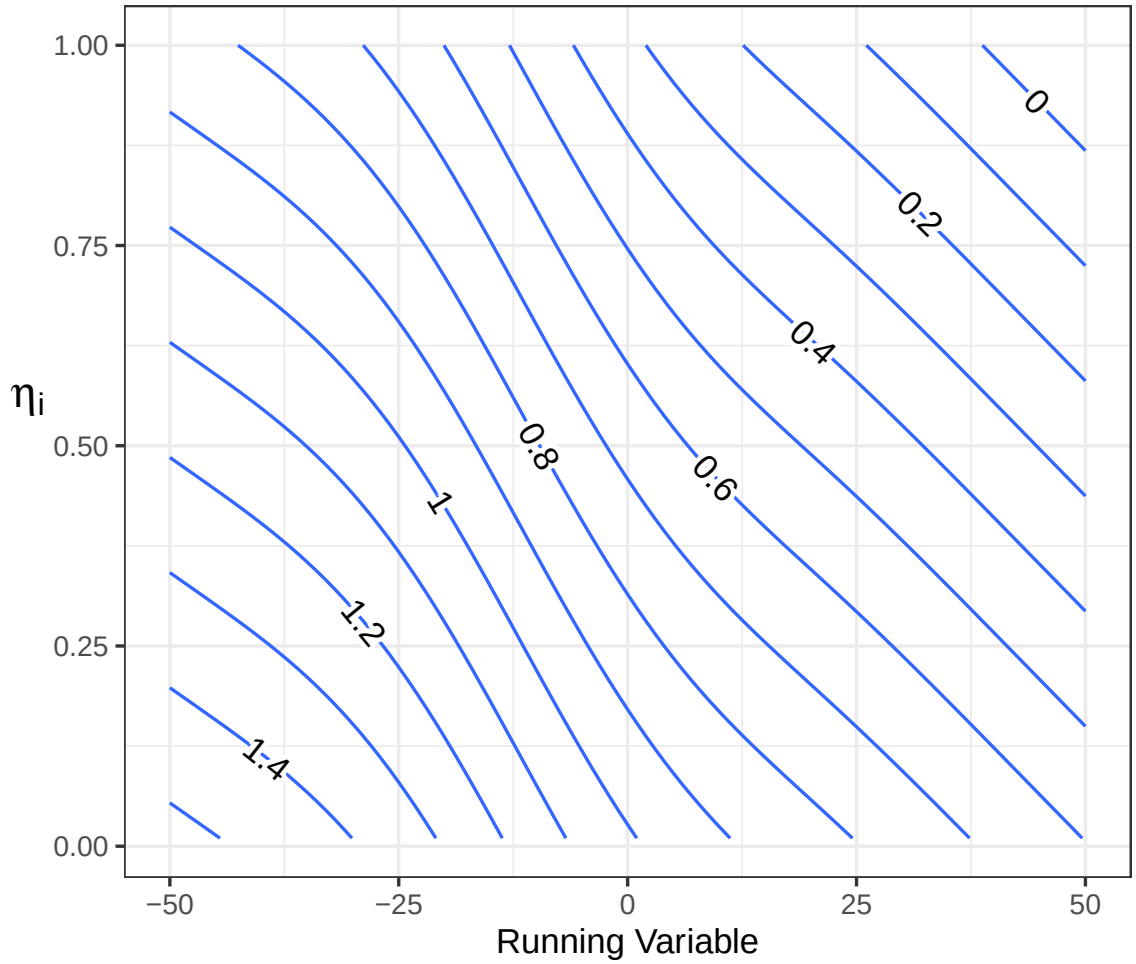
VI.A Graphs

Figure 1: Average Treatment Effect on the Treated



Note: The figure plots how the estimated the conditional average treatment effect on the treated varies with the running variable. Specifically, the solid line shows the estimated $\hat{\mathbb{E}}[\tau_i|Z_i = Z, T_i = 1]$ and the dashed lines indicated the 95% confidence interval, estimated via a Bayesian bootstrap with school-level clustering.

Figure 2: Estimates of $\tau(\eta, Z)$



Note: The figure illustrates how $\hat{\tau}(\eta, Z) = \mathbb{E}[\tau_i | \eta_i = \eta, Z_i = Z]$ varies with both Z and η. Each line corresponds to a set of (η, Z) values with the same value of $\hat{\tau}(\eta, Z)$. Roughly speaking, η_i is a latent variable that serves as a measure of how likely an individual is to enroll in the treatment; individuals' with low values of η_i are more likely to enroll than individuals with high values and so it is sometimes referred to as the "latent cost" of enrolling. See Section II.A for the formal definition.

VI.B Tables

Table 1: Average Effect with Different Treatment Assignments

	Optimal Assignment	Realized Assignment (ATT)	Local Effect (LATE)	Random Assignment (ATE)	Program Expansion (ATC)
Grade 4	1.36	1.11	0.90	0.50	0.42
	(0.42)	(0.11)	(0.12)	(0.47)	(0.53)
Grade 5	0.84	0.74	0.62	0.37	0.31
	(0.15)	(0.11)	(0.11)	(0.27)	(0.33)
Grade 6	0.69	0.61	0.53	0.07	−0.00
	(0.11)	(0.09)	(0.11)	(0.35)	(0.39)
Grade 7	0.53	0.45	0.36	0.01	−0.04
	(0.11)	(0.10)	(0.09)	(0.37)	(0.41)
Grade 8	0.61	0.47	0.39	0.12	0.08
	(0.22)	(0.11)	(0.08)	(0.44)	(0.48)

Note: Standard errors, generated via a Bayesian bootstrap procedure, are shown in parentheses. Optimal Assignment keeps the fraction of individuals treated fixed, but assigns the individuals with the highest treatment effects to the treatment. Realized Assignment is the average treatment effect of the realized assignment, which corresponds to the average treatment on the treated (ATT). Local Effect corresponds to the effect of the program on compliers at the treatment threshold (LATE). Random Assignment is the average treatment effect if treatment was assigned randomly, which corresponds to the overall average treatment on the treated (ATE). Program Expansion is the average treatment effect if treatment expanded to the individuals not currently receiving the treatment and corresponds to the average treatment on the controls (ATC).

References

Abadie, Alberto and Matias D. Cattaneo, "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 2018, *10* (1), 465–503.

Angrist, Joshua D. and Miikka Rokkanen, "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff," *Journal of the American Statistical Association*, 2015, *110* (512), 1331–1344.

Bertanha, Marinho and Guido W. Imbens, "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 2020, *38* (3), 593–612.

Boon, Michele Hilton, Peter Craig, Hilary Thomson, Mhairi Campbell, and Laurence Moore, "Regression Discontinuity Designs in Health: A Systematic Review," *Epidemiology*, 2021, *32* (1), 87–93.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes, "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, November 2020, *110* (11), 3634–60.

Cattaneo, Matias D., Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare, "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs," *Journal of the American Statistical Association*, 2021, *116* (536), 1941–1952.

Cook, Thomas D., ""Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics," *Journal of Econometrics*, 2008, *142* (2), 636–654. The regression discontinuity design: Theory and applications.

Dong, Yingying and Arthur Lewbel, "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics*, 2015, *97* (5), 1081–1092.

Figlio, David and Umut Özek, "An extra year to learn English? Early grade retention and the human capital development of English learners," *Journal of Public Economics*, 2020, *186*, 104184.

Freyberger, Joachim and Matthew A. Masten, "A practical guide to compact infinite dimensional parameter spaces," *Econometric Reviews*, 2019, *38* (9), 979–1006.

Greene, Jay and Marcus Winters, "Revisiting grade retention: An evaluation of Florida's test-based promotion policy," *Education Finance and Policy*, 2007, *2* (4), 319–340.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics, 2009.

Heckman, James J., "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 2010, *48* (2), 356–398.

_ **and Edward J. Vytlacil**, "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, 2007, chapter 70, pp. 4779–4874.

_ **and** _ , "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treat- ment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Vol. 6, Elsevier, 2007, chapter 71, pp. 4785–5143.

_ **and Edward Vytlacil**, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 1999, *96* (8), 4730–4734.

_ **and** _ , "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

Hwang, NaYoung and Cory Koedel, "Holding back to move forward: The effects of retention in the third grade on student outcomes," 2022.

Imbens, Guido W. and Thomas Lemieux, "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 2008, *142* (2), 615–635. The regression discontinuity design: Theory and applications.

Kline, Patrick and Christopher R. Walters, "On Heckits, LATE, and Numberican Equivalence," *Econometrica*, March 2019, *87* (2), 677–696.

Kowalski, Amanda, "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," *Review of Economics and Statistics*, forthcoming.

Licalsi, Christina, Umut Özek, and David Figlio, "The uneven implementation of universal school policies: Maternal education and Florida's mandatory grade retention," *Education Finance and Policy*, 2019, *14* (3), 383–413.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky, "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *Econometrica*, 2018, *86* (5), 1589–1619.

Mulhern, Christine, Isaac M. Opper, Fatih Unlu, Brian Phillips, and Julie Edmunds, "Dual Method of Dual Enrollment: Combining empirical approaches to estimate the impacts of taking college courses in high school on educational attainment," 2023.

Mumma, Kirsten and Marcus Winters, "The effect of retention under Mississippi's test-based promotion policy," 2023.

Newey, Whitney K. and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in R.F. Engle and D.L. McFadden, eds., *Handbook of Econometrics, Volume IV*, Elsevier Science, 1994.

Opper, Isaac M., "From LATE to ATE: A Bayesian Approach," 2023.

Özek, Umut, "Hold back to move forward? Early grade retention and student misbehavior," *Education Finance and Policy*, 2015, *10* (3), 350–377.

Rasmussen, Carl Edward and Chrisopher K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.

Schwerdt, Guido, Martin West, and Marcus Winters, "The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida," *Journal of Public Economics*, 2017, *152*, 154–169.

Winters, Marcus and Jay Greene, "The medium-run effects of Florida's test-based promotion policy," *Education Finance and Policy*, 2012, *7* (3), 305–330.

Wuepper, David and Robert Finger, "Regression discontinuity designs in agricultural and environmental economics," *European Review of Agricultural Economics*, 10 2023, *50* (1), 1–28.

A Proofs

A.A Main Proofs

Proposition 1. *Let $\hat{\tau}_{GRDD}$ be the estimate generated by the Global Regression Discontinuity Design, as defined above. Then given Assumptions 1 - 6, there exists τ^*_{GRDD} such that $\hat{\tau}_{GRDD} \xrightarrow{p} \tau^*_{GRDD}$.*

Proof. Our approach is to first show that there exists $k^*_{GRDD} \in \mathcal{K}_R$ such that that $\hat{k} \xrightarrow{p} k^*_{GRDD}$ and then appeal to the fact that T is continuous to conclude that $\hat{\tau}_{GRDD} = T(\hat{k}) \xrightarrow{p} T(k^*_{GRDD}) \equiv \tau^*_{GRDD}$.

To show that there exists such a k^*_{GRDD}, we first add a bit of notation. We let:

$$\hat{Q}_n(k) = \frac{1}{n}\left\{ \sum_{\forall i} \Big(Y_i - (1-T_i)\cdot k_0\big(\hat{\nu}(Z_i), Z_i\big) - T_i \cdot k_1\big(\hat{\nu}(Z_i), Z_i\big)\Big)^2 + J_m(k) \right\} \quad (24)$$

and so $\hat{k}$ minimizes $\hat{Q}_n$ subject to the constraint that $\hat{k} \in \mathcal{K}_R$. Similarly, we let $Q_0(k)$ be defined as:

$$Q_0(k) = \mathbb{E}\left[\Big(Y_i - (1-T_i)\cdot k_0\big(\nu(Z_i), Z_i\big) - T_i \cdot k_1\big(\nu(Z_i), Z_i\big)\Big)^2 \right] \quad (25)$$

We then show that four assumptions in Theorem 2.1 of Newey and McFadden (1994) hold, i.e., that: (i) $Q_0(k)$ has a unique minimum among $k \in \mathcal{K}_R$; (ii) $\mathcal{K}_R$ is compact; (iii) $Q_0(k)$ is continuous; and (iv) $sup_{k\in\mathcal{K}_R}|\hat{Q}_n(k) - Q_0(k)| \xrightarrow{p} 0$. Since the four assumptions hold, we can then conclude that $\hat{k} \xrightarrow{p} \hat{k}^*$ for the unique $\hat{k}^* \in \mathcal{K}_R$ that minimizes $Q_0(k)$.

The proofs that conditions (iii) is straightforward to show and condition (ii) follows from Assumption 6. Furthermore, from Lemma 1 and the law of large number, we get that condition (iv) holds.

The only condition whose proof is unique to the this context is (i), i.e., that $Q_0(k)$ has a unique minimum when we restrict the possible functions to the set $\mathcal{K}_R$. We show in Lemma 2 that there exists a single $k \in \mathcal{K}_R$ in which:

$$\mathbb{E}[Y_i|Z_i, \eta, T_i] = \big(1-T_i\big)\cdot\big(\alpha^*\eta + \gamma^*(Z)\big) - T_i \cdot \big(\beta^*\eta + \delta^*(Z)\big) \quad (26)$$

for all observed points Z. This then is clearly the k that minimizes $Q_0(k)$ subject to

the constraint that $k \in \mathcal{K}_R$, i.e., is k^*_{GRDD}.

Finally, while not technically required for this proof, we can show that the T operator is the correct operator by noting that:

$$\frac{\partial k_0^*(\eta, Z)}{\partial \eta} = \frac{\partial}{\partial \eta}\left(\frac{1}{1-\eta}\int_\eta^1 \mu^*(\tilde{\eta}, Z)d\tilde{\eta}\right) \tag{27}$$

$$= \frac{1}{1-\eta}\left(k_0^*(\eta, Z) - \mu^*(\eta, Z)\right) \tag{28}$$

and

$$\frac{\partial k_1^*(\eta, Z)}{\partial \eta} = \frac{\partial}{\partial \eta}\left(\frac{1}{\eta}\int_0^\eta \mu^*(\tilde{\eta}, Z) + \tau^*(\tilde{\eta}, Z)d\tilde{\eta}\right) \tag{29}$$

$$= \frac{1}{\eta}\left(\mu^*(\tilde{\eta}, Z) + \tau^*(\tilde{\eta}, Z) - k_1^*(\eta, Z)\right) \tag{30}$$

from which we can do some algebra to get the result.

□

Proposition 2. *The estimated effect on the set of compliers at the Z^* converges to the true effect on that set, i.e.:*

$$\frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*)d\eta = \frac{1}{p_h - p_l}\int_{p_l}^{p_h} \tau^*(\eta, Z^*)d\eta \tag{11}$$

Proof. We start by noting that:

$$\int_{p_l}^{p_h} \tau^*(\eta, Z^*)d\eta = \Big(p_h k_1^*(p_h, Z^*) - p_l k_1^*(p_l, Z^*)\Big) - \Big((1-p_l)k_0^*(p_l, Z^*) - (1-p_h)k_0^*(p_l, Z^*)\Big)$$

and we can similarly write $\int_{p_l}^{p_h} \tau^*_{GRDD}(\eta, Z^*)d\eta$ as an expression of the k^*_{GRDD} observations as (p_h, Z^*) and (p_l, Z^*).

From the proof of Theorem 1, however, it is clear that k^*_{GRDD} equals k^* at every observed point, i.e., at every point $(\nu(Z), Z)$. Since both (p_h, Z^*) and (p_l, Z^*) are observed points - by definition of p_h and p_l in Assumption 4 – the theorem follows. □

Remark 1. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and $\tau^*_{obs}(Z)$ be the estimate generated from*

the traditional observational study, as defined in Equation (12). We then have:

$$\tau^*_{GRDD}(Z) = \tau^*_{obs}(Z) - b \tag{14}$$

where b is a measure of the bias in the observational estimates. Specifically, we have:

$$b = \xi_h \cdot \left(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\right) + \xi_l \cdot \left(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\right) \tag{15}$$

where $\xi_i \in \mathbb{R}$ *is a function of* p_h, p_l *and* $\nu(Z)$*, and* $\tau^*_{obs}(Z^*_h) = \lim_{Z \downarrow Z^*} \tau^*_{obs}(Z)$ *and* $\tau^*_{obs}(Z^*_l) = \lim_{Z \uparrow Z^*} \tau^*_{obs}(Z)$*.*

Proof. We start by rearranging the traditional RD estimate to show that:

$$\begin{aligned}
\tau^*_{RDD} &= \tau^*_{obs}(Z^*_l) + \frac{k^*_0(p_h, Z^*) - k^*_0(p_l, Z^*)}{p_h - p_l} \cdot (1 - p_h) + \frac{k^*_1(p_h, Z^*) - k^*_1(p_l, Z^*)}{p_h - p_l} \cdot p_h \\
&= \tau^*_{obs}(Z^*_h) + \frac{k^*_0(p_h, Z^*) - k^*_0(p_l, Z^*)}{p_h - p_l} \cdot (1 - p_l) + \frac{k^*_1(p_h, Z^*) - k^*_1(p_l, Z^*)}{p_h - p_l} \cdot p_l
\end{aligned}$$

where k^* are the true conditional moments at the specified points, which are observed by the assumptions regarding the regression discontinuity design.

We next note that in the global regression discontinuity design we restrict the functional form of the estimated moments to be of the form: $k_0(\nu(Z), Z) = \alpha\nu(Z) + \gamma(Z)$ and $k_1(\nu(Z), Z) = \beta\nu(Z) + \delta(Z)$. Thus, we get that:

$$\begin{aligned}
&\tau^*_{GRDD}(Z) = k_1(1, Z) - k_0(0, Z) \\
&= \Big(k_1(1, Z) - k_1(\nu(Z), Z)\Big) + \Big(k_1(\nu(Z), Z) - k_0(\nu(Z), Z)\Big) + \Big(k_0(\nu(Z), Z) - k_0(0, Z)\Big) \\
&= \big(1 - \nu(Z)\big) \cdot \beta + \tau^*_{obs}(Z) + \nu(Z) \cdot \alpha
\end{aligned}$$

We then connect the traditional RDD and the Global RDD estimate by noting that the GRDD estimation implies that:

$$\alpha = \frac{k^*_0(p_h, Z^*) - k^*_0(p_l, Z^*)}{p_h - p_l} \qquad \text{and} \qquad \beta = \frac{k^*_1(p_h, Z^*) - k^*_1(p_l, Z^*)}{p_h - p_l} \tag{31}$$

Thus, we get that

$$\begin{bmatrix} 1 - p_h & p_h \\ 1 - p_l & p_l \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \tau^*_{RDD} - \tau^*_{obs}(Z^*_l) \\ \tau^*_{RDD} - \tau^*_{obs}(Z^*_h) \end{bmatrix} \tag{32}$$

□

Solving for α and β and then plugging those into the equation for $\tau^*_{GRDD}(Z)$, we finally get that:

$$\tau^*_{GRDD}(Z) = \tau^*_{obs}(Z) + \left(\frac{p_h \cdot \nu(Z) - (1-p_h)\cdot(1-\nu(Z))}{p_h - p_l}\right)\cdot\Big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_h)\Big) + \left(\frac{(1-p_l)\cdot(1-\nu(Z)) - p_l\cdot\nu(Z)}{p_h - p_l}\right)\cdot\Big(\tau^*_{RDD} - \tau^*_{obs}(Z^*_l)\Big)$$

Remark 2. *Let τ^*_{GRDD} be the estimate generated by the Global Regression Discontinuity Design, as defined in Section II.B, and τ^*_{RDD} be the estimate generated from the traditional regression discontinuity design, as defined in Equation (13). We then have:*

$$\tau^*_{GRDD}(Z) = \tau^*_{RDD} + l_C + l_Z \tag{16}$$

where l_C adjusts for fact that τ^*_{RDD} is local to the set of compliers and l_Z adjusts for the fact that τ^*_{RDD} is local to the discontinuity. Specifically, we have that:

$$l_C = \big[\beta^*\cdot(1-p_h) + k^*_1(p_h, Z^*)\big] - \big[k^*_0(p_l, Z^*) - \alpha^*\cdot p_l\big] - \tau^*_{RDD} \tag{17}$$

$$l_Z = \tau^*_{obs}(Z) - \tau^*_{obs}(Z^*_h) + (\alpha^* - \beta^*)\cdot\Big(\nu(Z) - p_h\Big) \tag{18}$$

where $\alpha^* = \frac{k^*_0(p_h,Z^*) - k^*_0(p_l,Z^*)}{p_h - p_l}$, $\beta^* = \frac{k^*_1(p_h,Z^*) - k^*_1(p_l,Z^*)}{p_h - p_l}$.

Proof. From the previous theorem, we can write that:

$$\tau^*_{GRDD}(Z) = \tau^*_{obs}(Z) + \beta^*\cdot\big(1-\nu(Z)\big) + \alpha^*\cdot\nu(Z) \tag{33}$$

where α^* and β^* are defined in the statement of the theorem. We then note that by definition:

$$\tau^*_{obs}(Z) = k^*_1\big(\nu(Z), Z\big) - k^*_0\big(\nu(Z), Z\big) \tag{34}$$

Using the fact that $k^*_0(p_l, Z^*) - \alpha^*\cdot p_l = k^*_0(p_h, Z^*) - \alpha^*\cdot p_h$, we get that:

$$\tau^*_{RDD} + l_c = \tau^*_{obs}(Z_h) + \beta^*\cdot\big(1-p_h\big) + \alpha^*\cdot p_h \tag{35}$$

and so:

$$\tau^*_{GRDD}(Z) = \tau^*_{RDD} + l_c + \Big[\tau^*_{obs}(Z) + \beta^* \cdot \big(1 - \nu(Z)\big) + \alpha^* \cdot \nu(Z)\Big] - \Big[\tau^*_{obs}(Z_h) + \beta^* \cdot \big(1 - p_h\big) + \alpha^* \cdot p_h\Big] \tag{36}$$

$$= \tau^*_{RDD} + l_c + \Big[\tau^*_{obs}(Z) - \tau^*_{obs}(Z_h) + \big(\alpha^* - \beta^*\big) \cdot \big(\nu(Z) - p_h\big)\Big] \tag{37}$$

□

Proposition 3. *Suppose that $k^* \in \mathcal{K}_R$. Then the estimated MTE function converges to the true MTE function, i.e., $\tau^*_{GRDD} = \tau^*$.*

Proof. This follows directly from the fact that k^*_{GRDD} is the only $k \in \mathcal{K}_R$ such that $k^*_{GRDD}\big(\nu(Z), Z\big) = k^*\big(\nu(Z), Z\big)$, i.e., from Theorem 1 and Lemma 2. □

Proposition 4. *Define $\mathcal{K}_D$ to be the k functions that are consistent with the true conditional moments, i.e.,*

$$\mathcal{K}_D = \Big\{k : \mathbb{E}[Y_i|\nu(Z_i), Z_i, T_i] = (1-T_i)\cdot k_0\big(\nu(Z_i), Z_i\big) + T_i \cdot k_1\big(\nu(Z_i), Z_i\big) \text{ for all } z_i \in \mathbf{Z}\Big\}$$

Then $\tau^*_{GRDD} = T(k^*_{GRDD})$ where k^*_{GRDD} is the minimum element of the preordered set $\big(\mathcal{K}_D, \preceq\big)$ when the preorder $\preceq$ is defined as the lexicographic order on $c(k)$, defined above.

Proof. From the proof of Theorem 1, we get that $k^*_{GRDD} \in \mathcal{K}_D$. We therefore only need to show that any other $k \in \mathcal{K}_D$ is such that $c(k^*_{GRDD}) \preceq c(k)$. This again follows directly from Theorem 1, which shows that k^*_{GRDD} is the only function that is both separable and linear in η and is in $\mathcal{K}_D$. Thus, any other $k \in \mathcal{K}_D$ has either $c_{11}(k) > c_{11}(k^*_{GRDD})$ or $c_{11}(k) = c_{11}(k^*_{GRDD})$ and $c_{12}(k) > c_{12}(k^*_{GRDD})$, which implies that $c(k^*_{GRDD}) \preceq c(k)$. □

Proposition 5. *We say that the estimator $\hat{\tau}_a$ **dominates** $\hat{\tau}_b$ iff $\mathscr{L}_C(\hat{\tau}_a) < \mathscr{L}_C(\hat{\tau}_b)$ for all C. Then:*

1. *The global RD design dominates the traditional observational study.*
2. *If $\nu(Z) = p_h$ above the cutoff and $\nu(Z) = p_l$ below the cutoff, the global RD design dominates the traditional fuzzy regression discontinuity design.*
3. *No estimator dominates the global RD design.*

Proof. For the proof, we will assume that $\omega(\eta, Z) = 1$, e.g., we are interested in the ATE and Z is uniformly distributed, but this is just to simplify notation and the results extend to other choices of ω.

For the first part, we start by noting that at any Z, we can re-write $\int_0^1 \tau^*_{GRDD}(\eta, Z)d\eta = \alpha\nu(Z) + \beta(1-\nu(Z)) + k_1^*(\nu(Z), Z) - k_0^*(\nu(Z), Z)$, where α and β are defined as in Theorem 2. Notably, we can use the true conditional moment conditions $k_i^*(\nu(Z), Z)$ since those are the conditional moments that are observe; see proof of Theorem 1. Since $\int_0^1 \tau^*(\eta, Z)d\eta = k_1^*(1, Z) - k_0^*(0, Z)$, we can therefore re-write the loss function as:

$$l_k(\tau_{GRDD}, \tau^*) = \left(\int \alpha(\nu(Z) + \beta\big(1 - \nu(Z)\big) - \Delta k^*(\nu(Z), Z)dZ \right)^2 \tag{38}$$

where

$$\Delta k^*(Z) \equiv \Big(k_1^*(1, Z) - k_1^*(\nu(Z), Z)\Big) - \Big(k_0^*(0, Z) - k_0^*(\nu(Z), Z)\Big) \tag{39}$$

. We can similarly write the loss function when using τ_{obs} as:

$$l_k(\tau_{obs}, \tau^*) = \left(\int \Delta k^*(Z)dZ \right)^2 \tag{40}$$

Now, rather than considering $\mathbb{E}[l_k(\tau^*_{GRDD}, \tau^*)]$, we consider the conditional expectation when conditioning on the four observed moments at the discontinuity, i.e., $k_1^*(p_h, Z^*)$, $k_0^*(p_h, Z^*)$, $k_1^*(p_l, Z^*)$, and $k_0^*(p_l, Z^*)$; we will refer to the vector of all four of these moments as $\mathbf{k}^*$. Note that when conditioning on these moments, both α and β are deterministic. From the modified GP, we therefore get that

$$\begin{bmatrix} k_0^*(0, Z) \\ k_0^*(\nu(Z), Z) \\ k_1^*(1, Z) \\ k_1^*(\nu(Z), Z) \end{bmatrix} \sim N\left(\begin{bmatrix} k_0^*(p_l, Z^*) - \alpha p_l \\ k_0^*(p_l, Z^*) - \alpha p_l + \alpha\nu(Z) \\ k_0^*(p_l, Z^*) - (1 - p_l)\beta \\ k_0^*(p_l, Z^*) - (1 - p_l)\beta + \beta(1 - \nu(Z)) \end{bmatrix}, \Sigma_C(Z) \right) \tag{41}$$

where $\Sigma_C(Z)$ is some positive definite covariance matrix that depends on the assumed covariance function C. See, for example, Equations (2.41) and (2.42) in Rasmussen and Williams (2006) and note that when conditioning only on $\mathbf{k}^*$ the linear model perfectly predicts the data. This means that, in their notation, $\mathbf{y} = H^T\overline{\beta}$, which

implies that $\overline{\mathbf{g}}(X_*) = H_*^T\overline{\beta}$ and that weights K_y^{-1} in the formula for the limiting $\overline{\beta}$ do not matter.

Thus, we get that:

$$\Delta k^*(Z)|\mathbf{k}^* \sim N\Big(\alpha(\nu(Z) + \beta\big(1-\nu(Z)\big), \sigma_C^2(Z)\Big) \tag{42}$$

where $\sigma_C^2(Z)$ depends on the assumed covariance function C.

We can extend this to get that for any finite number of points, we have that:

$$\Delta k^*(Z_i)|\mathbf{k}^* \sim N\Big(\alpha(\nu(Z_i) + \beta\big(1-\nu(Z_i)\big), \Sigma_C\Big) \tag{43}$$

where Σ_C is a (potentially large) covariance matrix.

Thus, we can conclude that:

$$\mathbb{E}\Big[l_k(\tau_{GRDD}, \tau^*)|\mathbf{k}^*\Big] = \mathbb{E}\left[\left(\int \alpha(\nu(Z) + \beta\big(1-\nu(Z)\big) - \Delta k^*(\nu(Z), Z)dZ\right)^2 \middle| \mathbf{k}^*\right] \tag{44}$$

$$= \lim_{N\to\infty} \mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N-1} \alpha(\nu(Z_j) + \beta\big(1-\nu(Z_j)\big) - \Delta k^*(\nu(Z_j), Z_j)\right)^2\right] \tag{45}$$

$$= \lim_{N\to\infty} \frac{\mathbf{1}\Sigma_C\mathbf{1}'}{N^2} \tag{46}$$

and

$$\mathbb{E}\Big[l_k(\tau_{obs}, \tau^*)|\mathbf{k}^*\Big] = \mathbb{E}\left[\left(\int \Delta k^*(\nu(Z), Z)dZ\right)^2 \middle| \mathbf{k}^*\right] \tag{47}$$

$$= \lim_{N\to\infty} \mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N-1} \Delta k^*(\nu(Z_j), Z_j)\right)^2 \middle| \mathbf{k}^*\right] \tag{48}$$

$$= \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N-1} \Big(\alpha(\nu(Z_j) + \beta\big(1-\nu(Z_j)\big)\Big)^2 + \frac{\mathbf{1}\Sigma_C\mathbf{1}'}{N^2} \tag{49}$$

where $Z_j = \underline{Z} + \frac{j}{N}\cdot(\overline{Z} - \underline{Z})$ and $\mathbf{1}$ is a vector of ones.

Thus, we have that $\mathbb{E}[l_k(\tau_{GRDD}, \tau^*)|\mathbf{k}^*] < \mathbb{E}[l_k(\tau_{obs}, \tau^*)|\mathbf{k}^*]$ for almost every $\mathbf{k}^*$. The one exception is when $\mathbf{k}^*$ is such that $\alpha = \beta = 0$; however, this is a zero-

probability event under our assumed GP. It is must therefore the case that:

$$\mathscr{L}_C(\tau_{GRDD}) = \mathbb{E}\big[\mathbb{E}[l_k(\tau_{GRDD}, \tau^*)|\mathbf{k}^*]\big] < \mathbb{E}\big[\mathbb{E}[l_k(\tau_{GRDD}, \tau^*)|\mathbf{k}^*]\big] = \mathscr{L}_C(\tau_{obs}) \quad (50)$$

for every C.

We take a similar approach to show that τ_{GRDD} dominates τ_{RDD}. For this, we start by noting that:

$$\tau_{RDD}(Z) - \tau^*(Z) = \big(\tau_{RDD}(Z) - \tau_{GRDD}(Z^*)\big) + \big(\tau_{GRDD}(Z^*) - \tau^*(Z)\big) \quad (51)$$

where $\tau_{GRDD}(Z^*)$ is the GRDD effect estimate at Z^*, which is equal to both $\alpha p_h + \beta(1-p_h) + k_1^*(p_h, Z^*) - k_0^*(p_h, Z^*)$ and $\alpha p_l + \beta(1-p_l) + k_1^*(p_l, Z^*) - k_0^*(p_l, Z^*)$.

Again conditioning on $\mathbf{k}^*$, we can get that:

$$\mathbb{E}\big[\big(\tau_{RDD}(Z) - \tau_{GRDD}(Z^*)\big) \cdot \big(\tau_{GRDD}(Z^*) - \tau^*(Z)\big)\big] = \quad (52)$$

$$\mathbb{E}\Big[\mathbb{E}\big[\big(\tau_{RDD}(Z) - \tau_{GRDD}(Z^*)\big) \cdot \big(\tau_{GRDD}(Z^*) - \tau^*(Z)\big)|\mathbf{k}^*\big]\Big] = \quad (53)$$

$$\mathbb{E}\Big[\big(\tau_{RDD}(Z) - \tau_{GRDD}(Z^*)\big) \cdot \mathbb{E}\big[\big(\tau_{GRDD}(Z^*) - \tau^*(Z)\big)|\mathbf{k}^*\big]\Big] = 0 \quad (54)$$

which stems from the fact that conditional on $\mathbf{k}^*$, both $\tau_{GRDD}(Z^*)$ and $\tau_{RDD}(Z)$ are deterministic and – for the same reasons above – that $\mathbb{E}\big[\big(\tau_{GRDD}(Z^*) - \tau^*(Z)\big)|\mathbf{k}^*\big] = 0$ for all $\mathbf{k}^*$. Thus, we can need only to show that $\mathbb{V}\big(\int \tau_{GRDD}(Z^*) - \tau^*(Z)dZ\big) > \mathbb{V}\big(\int \tau_{GRDD}(Z) - \tau^*(Z)dZ\big)$ to conclude that τ_{GRDD} dominates τ_{RDD}.

For this, we start by noting that:

$$\mathbb{V}\Big(\int \tau_{GRDD}(Z) - \tau^*(Z)dZ\Big) = \mathbb{V}\bigg(\int \alpha\nu(Z) + \beta\big(1-\nu(Z)\big) + \Delta k^*(\nu(Z), Z)dZ\bigg) \quad (55)$$

$$= \mathbb{V}\bigg(\int \Delta k^*(\nu(Z), Z)\bigg) - \mathbb{V}\bigg(\int \alpha\nu(Z) + \beta\big(1-\nu(Z)\big)dZ\bigg) \quad (56)$$

because:

$$\mathbb{C}\Big(\int \alpha\nu(Z) + \beta\big(1-\nu(Z)dZ, \int \Delta k^*(\nu(Z), Z)dZ\Big) = -\mathbb{V}\Big(\int \alpha\nu(Z) + \beta\big(1-\nu(Z)\big)dZ\Big) \quad (57)$$

which in turn follows from the fact that $\mathbb{E}\big[\sum_{j=1}^{N-1} \Delta k^*(\nu(Z_j), Z_j)|\mathbf{k}^*\big] = -\alpha\nu(Z_j) -$

$\beta\big(1-\nu(Z_j)$.[13]

We then note that the analysis of $\mathbb{V}\big(\int \tau_{GRDD}(Z^*)-\tau^*(Z)dZ\big)$ is nearly identical. The only difference is that instead of using $\Big(k_1^*(1,Z)-k_1^*(\nu(Z),Z)\Big)-\Big(k_0^*(0,Z)-k_0^*(\nu(Z),Z)\Big)$ for $\Delta k^*(\nu(Z_j),Z_j)$, we use $\Big(k_1^*(1,Z)-k_1^*(\nu(Z^*),Z^*)\Big)-\Big(k_0^*(0,Z)-k_0^*(\nu(Z^*),Z^*)\Big)$. In particular, we have that – under the assumption that $\nu(Z)$ is constant above and below the cutoff – the second term in Equation (56) is the same for both approaches and so we need only compare:

$$\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z)\big)+\big(k_0^*(0,Z)-k_0^*(\nu(Z),Z)\big)dZ\Big) \tag{58}$$

to

$$\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z^*),Z^*)\big)+\big(k_0^*(0,Z)-k_0^*(\nu(Z^*),Z^*)\big)dZ\Big) \tag{59}$$

To simplify the exposition, we will focus on comparing

$$\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z)\big)dZ\Big) \tag{60}$$

to

$$\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z^*)\big)dZ\Big) \tag{61}$$

from which we are able to conclude the full result, since we have assumed that k_1^* and k_0^* are generated from two independent GPs.

The fact that:

$$\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z)\big)dZ\Big) < \mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z^*)\big)dZ\Big) \tag{62}$$

follows from the Lemma 3, since $\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z^*)\big)$ is equivalent to the case where $\xi_l=\xi_h=0$ and $\mathbb{V}\Big(\int \big(k_1^*(1,Z)-k_1^*(\nu(Z),Z)\big)dZ\Big)$ is equivalent to the case where $\xi_l=Z^*-\underline{Z}$ and $\xi_h=\overline{Z}-Z^*$.

Finally, to show that no estimator dominates τ_{GRDD} we consider the case in which C is only a function of the running variable, i.e., $Cov(\tau^*(\eta_1,Z_1),\tau^*(\eta_2,Z_2))=c(Z_1,Z_2)$ for some function c. Then τ^* is separable in Z and η and linear in η. From Theorem 3, we then get that $\tau_{GRDD}=\tau^*$ and so $\mathscr{L}_C(\tau_{GRDD})=0$. Thus, there cannot

[13]Notably, if $\mathbb{E}[Y|X]=-X$, we get that $\mathbb{E}[(Y+X)^2]=\mathbb{E}[Y^2+2X\cdot Y+X^2]=\mathbb{E}[Y^2]+2\mathbb{E}[X\mathbb{E}[Y|X]]+\mathbb{E}[X^2]=\mathbb{E}[Y^2]-2\mathbb{E}[X^2]+\mathbb{E}[X^2]=\mathbb{E}[Y^2]-\mathbb{E}[X^2]$.

be another τ_{alg} such that $\mathscr{L}_C(\tau_{alg}) < \mathscr{L}_C(\tau_{GRDD})$ for all C.

□

A.B Supporting Lemmas

Lemma 1. *Let $\hat{\nu}_{GRDD}$ be the estimate of $\nu(Z_i)$ obtained in the Step 1 of the Global Regression Discontinuity Design approach defined in Section II.B. Then under Assumptions 2, 4, and 5 of Section II.A we have that $\hat{\nu}_{GRDD} \xrightarrow{p} \nu^*$.*

Proof. To show that there exists such a ν^*, we first add a bit of notation. We let:

$$\hat{V}_n(\nu) = \frac{1}{n}\left\{ \sum_{\forall i} \Big(T_i - \nu(Z_i)\Big)\Big)^2 + J_\nu(\nu) \right\} \tag{63}$$

and so $\hat{\nu}$ minimizes $\hat{V}_n$ subject to the constraint that $\hat{\nu} \in \mathcal{V}$. Similarly, we let $V_0(\nu)$ be defined as:

$$V_0(\nu) = \mathbb{E}\left[\Big(T_i - \nu(Z_i)\Big)^2 \right] \tag{64}$$

From our assumption that $T_i = \mathbf{1}(\nu^*(Z_i) \geq \eta_i)$, our normalization that $\eta_i \sim U(0,1)$, and the assumption on V it follows that $sup_{\nu \in V}|\hat{V}_n(\nu) - V_0(\nu)| \xrightarrow{p} 0$ and that ν^* minimizes $V_0(\nu)$. It is also clear that $V_0(\nu)$ is continuous and Assumption 5 ensures that $\mathcal{V}$ is compact. In short, the assumptions in Theorem 2.1 of Newey and McFadden (1994) hold; thus, $\hat{\nu}_{GRDD} \xrightarrow{p} \nu^*$. □

Lemma 2. *There exists a unique $k \in \mathcal{K}_R$ such that $k\big(\nu(Z), Z\big) = k^*\big(\nu(Z), Z\big)$ for all Z.*

Proof. From the assumption that there exists a discontinuity at Z^*, we know that at Z^* we observe $k_j^*(\eta, Z^*)$ at precisely two points – $k_j^*(p_h, Z^*)$ and $k_j^*(p_l, Z^*)$ – for $j \in \{0,1\}$. There is therefore a single choice of α and β that goes through the observed points at Z^*; we denote those as α^* and β^*. Since Z^* is the only point where we observe multiple values of $k_1^*(\eta, Z)$ and $k_0^*(\eta, Z)$, we can then set $\gamma^*(Z) = \mathbb{E}[Y_i|\nu^*(Z), Z, T_i = 0] - \alpha^*\nu^*(Z)$ and $\delta^*(Z) = \mathbb{E}[Y_i|\nu^*(Z), Z, T_i = 1] - \beta^*\nu^*(Z)$. These choices ensure that:

$$\mathbb{E}[Y_i|Z_i, \eta, T_i] = \big(1 - T_i\big) \cdot \big(\alpha^*\eta + \gamma^*(Z)\big) - T_i \cdot \big(\beta^*\eta + \delta^*(Z)\big) \tag{65}$$

for all observed points. □

Lemma 3. *Define* $f(Z, \xi_l, \xi_h)$ *as follows:*

$$f(Z,\xi_l,\xi_h) = \begin{cases} Z^* - \xi_l & \text{if } Z \leq Z^* - \xi_l \\ Z & \text{if } Z > Z^* - \xi_l \text{ and } < Z^* + \xi_h \\ Z^* + \xi_h & \text{if } Z \geq Z^* + \xi_h \end{cases} \tag{66}$$

Then $\mathbb{V}\Big(\int k_1\big(1, Z\big) - k_1\big(\nu(Z), f(Z,\xi_l,\xi_h)\big) dZ \Big)$ *is decreasing in both* ξ_l *and* ξ_h.

Proof. To simplify notation, let $\Delta k_1(Z,\xi_l,\xi_h) = k_1\big(1,Z\big) - k_1\big(\nu(Z), f(Z,\xi_l,\xi_h)\big)$. We then note that:

$$\mathbb{V}\Big(\int \Delta k_1(Z,\xi_l,\xi_h) dZ \Big) = \int\int \mathbb{C}\Big(\Delta k_1(Z,\xi_l,\xi_h), \Delta k_1(\tilde{Z},\xi_l,\xi_h) \Big) dZ d\tilde{Z} \tag{67}$$

We will next consider what happens when ξ_l increases to $\xi_l + \epsilon$, for some small $\epsilon > 0$. Next, note that if both $Z > Z^* - \xi_l$ and $\tilde{Z} - \xi_l$, then increasing ξ_l will not impact $\Delta k_1(Z,\xi_l,\xi_h)$ or $\Delta k_1(\tilde{Z},\xi_l,\xi_h)$. Thus, we can write:

$$\mathbb{V}\Big(\int \Delta k_1(Z,\xi_l+\epsilon,\xi_h) dZ \Big) - \mathbb{V}\Big(\int \Delta k_1(Z,\xi_l,\xi_h) dZ \Big) = \tag{68}$$

$$\int_{\underline{Z}}^{Z^*-\xi_l} \int_{\underline{Z}}^{Z^*-\xi_l} \mathbb{C}\Big(\Delta k_1(Z,\xi_l+\epsilon,\xi_h), \Delta k_1(\tilde{Z},\xi_l+\epsilon,\xi_h) \Big) - \mathbb{C}\Big(\Delta k_1(Z,\xi_l,\xi_h), \Delta k_1(\tilde{Z},\xi_l,\xi_h) \Big) dZ d\tilde{Z} + \tag{69}$$

$$2\int_{Z^*-\xi_l}^{\overline{Z}} \int_{\underline{Z}}^{Z^*-\xi_l} \mathbb{C}\Big(\Delta k_1(Z,\xi_l+\epsilon,\xi_h), \Delta k_1(\tilde{Z},\xi_l+\epsilon,\xi_h) \Big) - \mathbb{C}\Big(\Delta k_1(Z,\xi_l,\xi_h), \Delta k_1(\tilde{Z},\xi_l,\xi_h) \Big) dZ d\tilde{Z} \tag{70}$$

Again, to slightly simplify notation we will consider the derivative as ξ_l increases. For the first term, consider any $Z, \tilde{Z} < \xi_l$. Then expanding the two terms of Δk_1 and taking the covariances, we get that:

$$\mathbb{C}\Big(\Delta k_1(Z,\xi_l+\epsilon,\xi_h), \Delta k_1(\tilde{Z},\xi_l+\epsilon,\xi_h) \Big) - \mathbb{C}\Big(\Delta k_1(Z,\xi_l,\xi_h), \Delta k_1(\tilde{Z},\xi_l,\xi_h) \Big) = \tag{71}$$

$$C'_{1,Z}(|Z - (Z^* - \xi_l)|) C_{1,\eta}(1,\nu(Z)) + C'_{1,Z}(|\tilde{Z} - (Z^* - \xi_l)|) C_{1,\eta}(1,\nu(\tilde{Z})) \tag{72}$$

$$< 0 \tag{73}$$

where the fact that it's less than zero stems from the assumption that $C_{1,Z}(|Z - Z'|)$ is decreasing in $|Z - Z'|$ and that $C_{1,\eta}(\eta, \eta') > 0$ for all η, η'.

For the second term, we consider any $Z < Z^* - \xi_l$ and $\tilde{Z} > Z^* - \xi_l$. In this case, we can again expand the two terms of Δk_1 and take the covariances to get that:

$$\mathbb{C}\Big(\Delta k_1(Z, \xi_l + \epsilon, \xi_h), \Delta k_1(\tilde{Z}, \xi_l + \epsilon, \xi_h)\Big) - \mathbb{C}\Big(\Delta k_1(Z, \xi_l, \xi_h), \Delta k_1(\tilde{Z}, \xi_l, \xi_h)\Big) = \quad (74)$$

$$C'_{1,Z}(|\tilde{Z} - (Z^* - \xi_l)|)C_{1,\eta}(1, \nu(Z)) - C'_{1,Z}(|\tilde{Z} - (Z^* - \xi_l)|)C_{1,\eta}(\nu(\tilde{Z}), \nu(Z)) = \quad (75)$$

$$C'_{1,Z}(|\tilde{Z} - (Z^* - \xi_l)|) \cdot \Big[C_{1,\eta}(1, \nu(Z)) - C_{1,\eta}(\nu(\tilde{Z}), \nu(Z))\Big] \quad (76)$$

This is negative as long as $C_{1,\eta}(\nu(\tilde{Z}), \nu(Z)) > C_{1,\eta}(1, \nu(Z))$. Note that when $\tilde{Z} < Z^*$, $\nu(\tilde{Z}) = \nu(Z)$ and so $C_{1,\eta}(\nu(\tilde{Z}), \nu(Z))$ is clearly greater than $C_{1,\eta}(1, \nu(Z))$. When $\tilde{Z} > Z^*$, in contrast, the question is whether $C_{1,\eta}(p_h, p_l)$ is greater than $C_{1,\eta}(1, p_h)$ which is less clear.

Note, however, that the negative component (i.e., Equation 72) gets more negative as the size of the direct effect (i.e., c_η) increases, while the size of the direct effect has no impact on the ambiguous component (i.e., Equation 76). Thus, if the direct effect is large enough, the negative term dominates and the variance is decreasing in ξ_l.

The proof that the result also holds for ξ_h follows similarly.

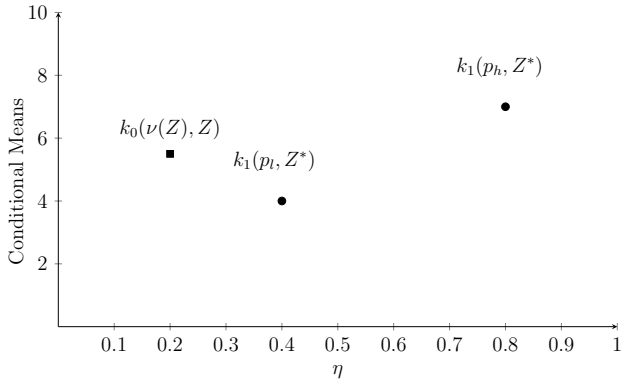
□

B Identification Intuition

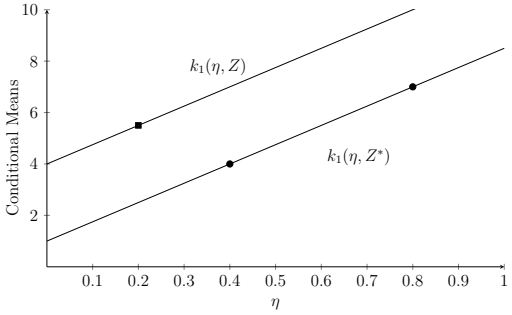
To highlight the identification challenge and convey the intuition behind our approach, consider the simplified example in which we focus only on the function k_1 and only observe the conditional averages at the discontinuity – $k_1(p_l, Z^*)$ and $k_1(p_h, Z^*)$ – and at a second point away from the discontinuity – here $k_1(\nu(Z), Z)$. These three points are illustrated in Figure 3a. While we focus on this simple example, the intuition would equally apply if we were aiming to estimate both k_1 and k_0 and we observe N points instead of three.

Of course, there are many functions $k_1(\eta, Z)$ that go through these three points, which suggests that the function is not fully identified. This identification challenge is mitigated if we assume – as we do in the paper – that $k_1(\eta, Z) = \alpha\eta + \gamma(Z)$ for some $\alpha \in \mathbb{R}$ and function $\gamma(Z)$. We then are able to identify α from the fact that

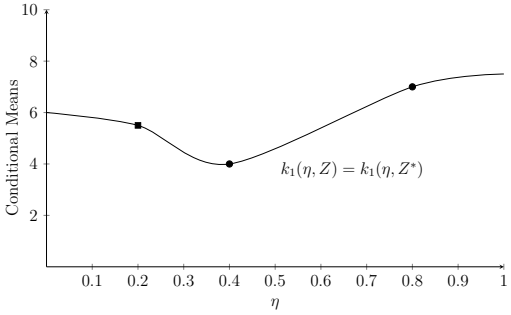
Figure 3: Identification Intuition



(a) Observed conditional moments



(b) Linear selection



(c) Nonlinear selection

we observe two points at Z^* and then can choose $\gamma(Z)$ to go through every other observed point. The result is illustrated in Figure 3b. In this case, we prioritize the fact that $k_1(\eta, Z)$ is separable and linear in η over ensuring that $\gamma(Z)$ is a smooth function of Z. This prioritization shows up in the lexicographic preferences defined in Section III.B.

From this example, it is clear that could define a similar estimator under a different prioritization scheme. At the opposite extreme, for example, one could allow k_1 to be a flexible function of η, but restrict the function to not vary based on Z. Figure 3c illustrates what the resulting function $k_1(\eta) = k_1(\eta, Z^*) = k_1(\eta, Z)$ might look like.

C Simulations

In additional to our empirical application, we conduct a simulation to compare the performance of the proposed approach (i.e., the global RD design) to the two most

natural alternatives – an observational study and a traditional RD design. The advantage of the simulation is that we can compare the estimated effect to the true effect. For our simulation, we assume that the true moments are generated according to the modified GP as defined in Section III.B. In particular, we assume that:

$$\mu(\eta, Z) \sim GP(0, \Sigma_{\mu,\eta}) + N(0,1) * \eta + GP(0, \Sigma_{\mu,Z}) \tag{77}$$

$$\tau(\eta, Z) \sim GP(0, \Sigma_{\tau,\eta}) + N(0,1) * \eta + GP(0, \Sigma_{\tau,Z}) \tag{78}$$

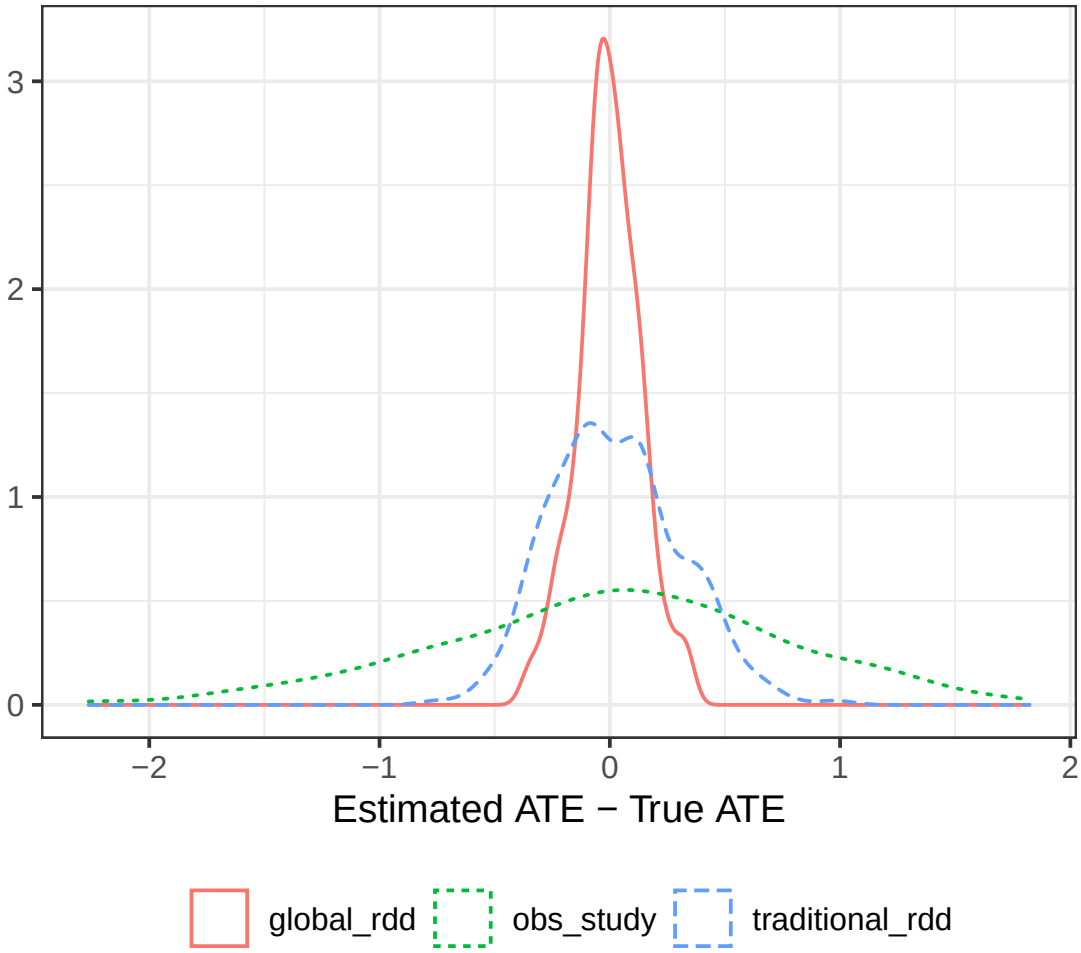
where $GP(0, \Sigma)$ corresponds to a Gaussian process with covariance Σ. For our covariance functions, we use a squared exponential with length scale 1 for $\Sigma_{\mu,\eta}$ and $\Sigma_{\tau,\eta}$ and length scale 2 for $\Sigma_{\mu,Z}$ and $\Sigma_{\tau,Z}$. We use an output variance of 1 for $\Sigma_{\mu,\eta}$ and $\Sigma_{\mu,Z}$ and of 0.5 for the other two GPs. We ignore statistical uncertainty by assuming that we observe the true conditional moments without error, although we only observe the points $\big(\nu(Z), Z\big)$, where:

$$\nu(Z) = \Phi\Big(-0.75 + 0.75 * Z + 0.75 * \mathbf{1}(Z > -.5)\Big) \tag{79}$$

where $\Phi(.)$ is the normal CDF and $\mathbf{1}$ represents the indicator function. Finally, we specify that $Z_i \sim U(-1, 1)$ and focus on the average treatment effect (ATE).

As can be seen in Figure 4, the proposed estimator outperforms the others two; this is consistent with the results of Theorem 5. Specifically, the mean-squared error of the global RD design is 0.02, while the mean-square error is 0.08 for the traditional RD design and 0.57 for the observational study.

Figure 4: Simulation Distributions



Note: The figure plots kernel densities showing have three estimators – the global RDD (*global_rdd*) proposed in this paper, as well as the observational study (*obs_study*) and a traditional RDD (*traditional_rdd*) formally defined in Section III.A – perform under a simulation. To create the figure, we substract the true ATE from the estimated ATE and then plotted kernel densities of this measure for each of the three estimators.